DOCTOR OF PHILOSOPHY

Argument revision and its role in dialogue

Snaith, Mark Ian

Award date:
2013

DOCTOR OF PHILOSOPHY

Argument revision and its role in dialogue

Mark Ian Snaith

2013

University of Dundee

Argument Revision and its Role in Dialogue

Mark Ian Snaith

Doctor of Philosophy

University of Dundee

Scotland

December 2012

Contents

Declarations **x**

1 Introduction **1**

1.1 Overview 1

1.2 Motivation 3

1.3 Research hypothesis 6

1.4 Thesis structure 7

2 Background **9**

2.1 Introduction 9

2.2 Argumentation 9

2.2.1 Argumentation Frameworks 10

2.2.2 Extensions, instantiations and implementations of Dung [1995] . 15

2.2.3 Defeasible argumentation 21

2.3 Belief revision 22

2.3.1 AGM Postulates 23

2.3.2 Epistemic entrenchment 25

2.4 Dialogue 26

2.4.1 Dishonesty in dialogue 27

2.4.2 Dialogue in multi-agent systems 28

2.5 Argumentation, belief revision and dialogue 28
2.5.1 Argumentation and dialogue 28
2.5.2 Argumentation and Belief Revision 29
2.5.3 Belief revision and dialogue 33
2.6 Summary and discussion 33

3 Logical preliminaries **37**
3.1 Introduction 37
3.2 The ASPIC$^+$ framework 37
3.3 Meta-argumentation 42
3.4 Summary 49

4 Dialogue framework: SPD **51**
4.1 Introduction 51
4.2 Personal and shared argumentation theories 52
4.2.1 Personal argumentation theory 52
4.2.2 Shared argumentation theory 54
4.3 Communication language 55
4.4 Commitment and commitment rules 57
4.4.1 Commitment store 57
4.4.2 Commitment rules 58
4.5 Structural rules 61
4.6 Running example 63
4.7 Summary 69

5 Argument Revision **71**
5.1 Introduction 71
5.2 Argument contraction and expansion 74
5.2.1 Principles of Argument Revision 74

5.2.2 Argument contraction 76
5.2.3 Argument expansion 76
5.3 Process of Argument Revision 77
5.3.1 Formula removal 79
5.3.2 Formula addition 80
5.3.3 Change graphs 82
5.4 Rule, preference and contrariness-based Argument Revision 84
5.4.1 Rule-based Argument Revision 84
5.4.2 Contrariness and preference-based Argument Revision 88
5.5 Properties of Argument Revision 90
5.5.1 Combining removal and addition 90
5.5.2 Structural properties 91
5.6 Measures of minimal change 92
5.7 Change graph example 99
5.8 Summary 103

6 Argument revision in dialogue **104**

6.1 Introduction 104
6.2 Commitment retraction 106
6.2.1 Stability adjustments 106
6.2.2 Choosing what to retract 108
6.3 Lying 112
6.3.1 Why lie? 113
6.3.2 Characterisation of lying in dialogue 115
6.3.3 Argument revision and dishonesty 116
6.3.4 Running example 118
6.4 Summary 122

7 Conclusions **124**

7.1 Summary 124

7.2 Contributions 126

7.2.1 Meta-level extensions to the ASPIC$^+$ framework 127

7.2.2 Revision without entrenchment 128

7.2.3 Decoupling dynamics from the system 129

7.3 Future work 129

7.3.1 Theoretical work 129

7.3.2 Applications 131

7.4 Validating the research hypothesis 132

7.5 Conclusions 133

List of Figures

1.1 Revision process for dishonesty 6

2.1 Framework for Example 2.2.1 11
2.2 Framework illustrating the problem with sceptical-preferred semantics . . 14

4.1 Framework from $\mathcal{PAT}_\alpha$ 66
4.2 Framework from $\mathcal{PAT}_\beta$ 67
4.3 Combined framework from $\mathcal{PAT}_\alpha$ and $\mathcal{PAT}_\beta$ 68

5.1 Framework using strict rule $p \rightarrow s$ with $\mathcal{R}_s$ closed under transposition . . 86
5.2 Framework using defeasible rule $p \Rightarrow s$ 87
5.3 Argumentation framework with all arguments "undecided" 89
5.4 Argumentation framework with $\mathcal{A}_2$ and $\mathcal{A}_3$ "in" and $\mathcal{A}_1$ "out" 89
5.5 Argumentation framework with $\mathcal{A}_5$ "out" 94
5.6 Argumentation framework with $\mathcal{A}_5$ "in" 95
5.7 Abstract framework from $\mathcal{AT}$ 101
5.8 Change graph for $\mathcal{AT} \dot{-} \{\mathcal{A}_6, \mathcal{A}_{10}\}$ 102

6.1 Dialogue move selection 105
6.2 Stability adjustment from [Walton and Krabbe, 1995] 107
6.3 Alternative stability adjustment 107

6.4 Change graph for $\mathcal{PAT}_\alpha \dot{-} \{\mathcal{A}_{13}\}$ 110
6.5 Abstract framework for the politician example 115
6.6 Dialogue move selection incorporating lying (initial steps omitted) 119
6.7 Framework from $\mathcal{PAT}_\alpha$, incorporating β's argument for d 120
6.8 Change graph for $\mathcal{PAT}_\alpha \dot{+} \{\mathcal{A}_{19}\}$ 121

List of Tables

2.1 AGM postulates for revision 24
2.2 AGM postulates for contraction 24
2.3 Justifications for **(EE2)**-**(EE5)** 26
2.4 Structural properties for a change operation, from [Cayrol et al., 2010] . . 32

4.1 Dialogue fragment 69

5.1 Structural properties for a change operation, from [Cayrol et al., 2010] . . 91
5.2 Possible initial removals in $\mathcal{AT}\dot{-}\{\mathcal{A}_6, \mathcal{A}_{10}\}$ 101
5.3 Possible initial removals in $\mathcal{AT}\dot{-}\{\mathcal{A}_6, \mathcal{A}_{10}\}$ 102

6.1 Dialogue fragment, from Chapter 4 109
6.2 Continuation of the dialogue 109
6.3 Outputs of the functions for measuring minimal change 111
6.4 Dialogue progression with α retracting 111
6.5 Outputs of the functions for measuring minimal change in $\mathcal{PAT}_\alpha\dot{+}\{\mathcal{A}_{19}\}$ 120
6.6 Continuation of dialogue when α lies 122

In memory of my father

Acknowledgements

Many people played a part in this thesis coming to fruition. First and foremost, I'm grateful to my supervisor, Prof. Chris Reed for his support and guidance throughout. Thanks go also to the other members of ARG:dundee, past and present, who offered comments and advice on work-in-progress.

I'd also like to thank my friends here in the School of Computing for offering what were at times necessary distractions that focused the mind. On many occasions I'd return from lunch or coffee and a find a problem had solved itself.

Last, but my no means least, I'd like to thank my family for their continued encouragement and support.

The research presented in this thesis was funded by the Engineering and Physical Sciences Research Council (EPSRC) of the UK government under grant number EP/G060347/1.

Declarations

Candidate's Declaration

I, Mark Snaith, hereby declare that I am the author of this thesis; that I have consulted all references cited; that I have done all the work recorded by this thesis; and that it has not been previously accepted for a degree.

Supervisor's Declaration

I, Chris Reed, hereby declare that I am the supervisor of the candidate, and that the conditions of the relevant Ordinance and Regulations have been fulfilled.

Abstract

In this thesis, a model for argument revision is presented, in terms of the expansion and contraction of a system of structured argumentation. At its core, the model uses the belief revision concept of minimal change, but without requiring a pre-determined entrenchment ordering to establish minimality.

In the first part of the thesis, a model for argument revision is defined and described. Specified in terms of the ASPIC$^+$ framework for argumentation, the model is divided into two main concepts: argument expansion, whose goal is to make certain arguments acceptable in the system, possibly by adding them; and argument contraction, whose goal is to make certain arguments unacceptable in the system, possibly by removing them. The goal of a revision process can be achieved in multiple different ways, thus a method of choosing which, based on measures of minimal change, is also specified.

The second part of the thesis demonstrates two applications of the model in the context of multi-agent dialogue. The first is used to assist a participant when faced with a need to update its commitment store during persuasion dialogue, while the second shows how a participant can use argument revision techniques to both assess and maintain a lie.

The main contributions of the thesis are twofold. First, the characterisation of a model for argument revision, based on established belief revision principles but with a key difference. The model for argument revision demonstrates how it is possible to use measurable effects on the system when determining minimal change instead of relying on

a pre-determined, qualitative entrenchment ordering.

Second, the thesis demonstrates two applications of argument revision in dialogue. The first is in assisting an agent in retracting a commitment that has been defeated, and for which it can offer no defence. When retracting a claim, the participant may also be required to retract other claims from which the defeated one is a consequence. Applying argument revision techniques allow the participant to reason about what constitutes a minimal set of retractions, in terms of current commitments and potential future communications in the dialogue.

The second dialogical application relates to the opposite of retraction; instead of choosing to retract an undefended claim, the participant could instead choose to lie in order to defend it. Argument revision allows the participant to not only assess whether or not lying is "minimal" (compared to retracting), but to also to maintain the lie, by using the measures of minimal change.

Overall, the thesis shows that not only is justifiable argument revision possible without relying on a pre-determined entrenchment ordering, it is also a powerful tool for participants in a dialogue, assisting with dialogue move selection.

Chapter 1

Introduction

1.1 Overview

Argumentation is a broad research field, with contributions from computer science, philosophy, discourse analysis and law. An emerging area in the field is the connection between argumentation and artificial intelligence, specifically the artificial modelling of human-style reasoning, using systems of argumentation with established theories of inference, conflict and evaluation. However, in addition to these, the modelling of further concepts are required; one such concept is dynamics — how the system reacts to change, in terms of gaining and losing information.

Belief revision is the study of belief dynamics, with one of the most prominent theories being the eponymous AGM theory [Alchourrón et al., 1985], which is based on a principle of *minimal change* — when updating a belief set, it should be done so with the smallest possible change on other beliefs. Research into connections between belief revision and argumentation has recently found new momentum [Falappa et al., 2009, 2012], with two broad areas emerging — the use of argumentation to assist the belief revision process (e.g. [Krümpelmann et al., 2012]), and the use of belief revision techniques to model the dynamics of argumentation systems (e.g. [Rotstein et al., 2008]).

In this thesis, a model for *Argument Revision* is specified in which the dynamics of an

existing system of structured argumentation are modelled through the application of the belief revision principle of minimal change. This model differs from existing approaches to specifying argument dynamics, such as that found in Rotstein et al.'s [2008] Argument Theory Change (ATC) by decoupling the model of dynamics and making it extrinsic to the system of argumentation to which it applies. ATC defines a new Dynamic Argumentation Framework, based on a Dung [1995] style abstract framework, which incorporates a sub-argument relation and a concept of warranting, which is used as part of the dynamics model. The model of argument revision presented here takes an existing system of structured argumentation and specifies revision operators and a determination of minimal change based on elements of the system. To assist with the specification of Argument Revision, a principled definition of meta-level extensions to the system is also provided, which offer wider advantages and applications beyond the modelling of dynamics.

The thesis also explores an application of the model as a strategic tool in multi-agent dialogue. A clear, but under-researched, link exists between belief revision and philosophical and computational models of dialogue [Girle, 1997, 2002]. Protocols based on the Walton and Krabbe [1995] dialogue typology provide each participant with a commitment store, and a record of statements they have uttered and/or conceded during the dialogue. This, by definition, is dynamic — statements are added and removed as the dialogue progresses. Statements are removed from a commitment store through retractions, which some protocols, such as RPD_0, mandating that when a statement is retracted, other statements of which the retracted statement is a consequence must also be retracted; this is known as a *stability adjustment*.

Selecting what, if any, statements to retract illustrates a connection between dialogue and belief revision — intuitively, a participant would retract the statements that, through applying some criteria, have a minimal impact on both existing commitments and also potential future commitments. The impact on existing commitments arises through introducing inconsistencies and/or removing statements that prevent a given statement from being inferred (under closure). The impact on potential future commitments comes from

the inability to communicate a statement of which a previously retracted statement is a consequence (for instance, if P is a consequence of Q and Q is retracted, P cannot then be stated).

A second application of *Argument Revision* in dialogue, dishonesty, is also explored. Being dishonest is a form of belief revision and, by extension *Argument Revision*, in that although dishonesty does not see a dialogue participant change their *actual* beliefs, they must perform a revision process (w.r.t. the source of the dishonesty) in order to determine an epistemic state they must *appear* to hold in order to avoid detection. Additionally, where multiple possible dishonest acts exist, minimal change is used to decide which to choose.

1.2 Motivation

Modelling the argument dynamics is important if we are to effectively model human reasoning in systems of argumentation. When a human is faced with new information, they (perhaps subconsciously) perform an assessment of the information to determine how it fits into their current beliefs and the effects thereof. Investigating the dynamics of belief forms the basis of a research field broadly described as *belief revision*. One of the most prominent and influential theories in belief revision is the AGM theory [Alchourrón et al., 1985], which describes operations, and the properties thereof, for three types of belief dynamic: **expansion**, where new beliefs are added to a belief set, with no consideration for consistency; **contraction**, where existing beliefs are removed; and **revision**, where new beliefs are added and old ones given up (where necessary) to maintain consistency. One of the guiding principles in all three processes is *minimal change* — when multiple methods of performing a process exist, choose the one that has the least impact on existing or remaining beliefs.

Given the prominence of Alchourrón et al.'s work, it is not surprising that it has formed the basis of existing research into connections between argumentation and belief revision

that focus on the modelling of argument dynamics. Rotstein et al. [2008] define a dynamic argumentation framework whose dynamics are modelled using AGM-style operators, while certain properties of change in abstract argumentation defined by Cayrol et al. [2010] are also inspired by AGM concepts.

The dynamic argumentation framework defined in Rotstein et al.'s work is based on Dung's abstract frameworks, but incorporates a model of dynamics, while Cayrol et al. examine change in Dung-style frameworks "as-is", i.e. they do not define a new or extended framework, but instead investigate the properties of changing a framework constructed according to the original definition of [Dung, 1995].

Both approaches have their advantages; defining a new argumentation framework whose dynamics are a part of the system means that the system intrinsically contains what is required to model change and hence determinations of, for instance, the effects of a change will be more readily available. Additionally, through the use of a sub-argument relation, the framework defined in ATC is also structured, providing a better model of dynamics and a clear link to natural arguments, which are inherently structured. On the other hand, developing a model of dynamics for an existing framework that is based on an assessment of the effects of different types of change decouples the model from the framework to which it applies. This has the distinct advantage of being able to assess the dynamics of an existing framework, with established properties, without needing to extend the framework and hence prove its properties still hold. Furthermore, existing extensions and applications of a framework could, in principle, adopt such a model without needing to extend a new version of the framework.

Combining the approaches of Rotstein et al. and Cayrol et al. provides a solution — taking an existing system of *structured* argumentation and modelling its dynamics based on the effects of making changes to the system, without extending the system to accommodate the model of dynamics, combines the advantages of both approaches. This is what Argument Revision does, resulting in a more general, applicable and expressive model for revising a system of argumentation.

Argument dynamics are also important in argument-based dialogues, if the participants in the dialogue are to fully understand the impact of conceding statements to their opponent(s), and/or retracting statements they have previously made. When statements are conceded or retracted, there is not only an impact on what has already been expressed in the dialogue, but also on potential future utterances. Consider as a high-level example a participant conceding $\neg P$ to an opponent, when they actually believe (but have not yet uttered) P; the concession means that if they wish to remain consistent, they cannot utter P later in the dialogue (unless the future course of the dialogue allows $\neg P$ to be retracted).

Another application of argument dynamics in dialogue occurs when a participant is open to dishonesty. Instead of conceding a statement to their opponent, a participant could offer a counter-argument that they don't actually believe. While dishonesty does not involve revision of actual beliefs, it requires the participant to present an epistemic state that (if the liar is to be dialogically successful) consistently accommodates the source of the dishonest utterance. This allows dishonesty to be considered a form of belief (or argument) revision, where beliefs are put through a revision process with respect to the dishonest statement to yield what the participant must externally appear to believe. This process is summarised in Figure 1.1. In the context of argument dynamics, the participant would need to assess, and be aware of, the implications of accommodating the argument at the source of the dishonesty in whatever framework is being used as the source of their beliefs.

Furthermore, the formal account of lying specified by Sakama et al. [2010] requires that dishonest acts be kept 'minimal" with respect to beliefs so as to reduce the chances of the dishonesty being exposed. This again results in a natural link to the AGM model of belief revision, which is guided by the notion of minimal change; a minimal dishonest act is the one that results in the minimal change between actual and externally-presentable beliefs.

Given these two applications, assessing dynamics can be considered a strategic tool in a dialogue, in that it allows a participant to carefully plan its next move, both in terms

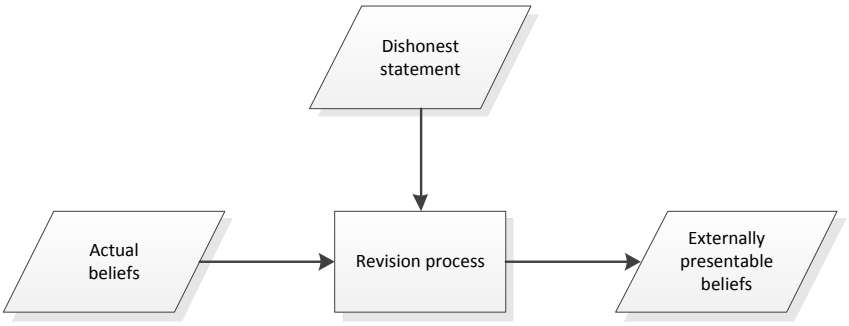

Figure 1.1: Revision process for dishonesty

of the locution and its content, through assessing the impact of that move on current and potential future commitments and, in the case of dishonest, external perceptions of its beliefs. These dialogical applications provide a further advantage of decoupling the model of dynamics from the framework for argumentation to which it is applied. Basing a dialogue on a framework that has a built-in model of dynamics would remove the strategic edge, since the model would be available to all participants. However, a decoupling allows each participant to individually determine whether or not to assess dynamics, and indeed each may employ different models, which yield different results. In effect, the model for dynamics is a component used to process the framework as opposed to an intrinsic part of it.

1.3 Research hypothesis

The research hypothesis upon which this thesis is based is as follows:

> *A model of Argument Revision can be developed that is applied to, but remains independent of, an existing system of argumentation, which in turn can be deployed as a strategic tool in multi-agent dialogue.*

The system of argumentation that will be used in answering this hypothesis is the ASPIC$^+$ framework characterised by Prakken [2010]. ASPIC$^+$ instantiates the abstract

approach to argumentation specified by Dung [1995] by combining the work of Pollock [1987] on defeasible reasoning with the work of Vreeswijk [1997] on the structure of arguments.

ASPIC$^+$ is chosen for several reasons:

1. It is a system of structured argumentation (as opposed to purely abstract) with certain proven properties, which provides the ability to explore the effect of dynamics on the structure of arguments, and the properties of the system as a whole.

2. The use of two types of inference rule (strict and defeasible, based on Pollock's [1987] theory of defeasible reasoning) presents an opportunity to investigate rule-based argument dynamics (i.e. revision of arguments based on the rules they apply).

3. Being based on Dung's theory, it takes advantage of existing, established acceptability semantics for evaluation. This allows for investigation into links between dynamics and acceptability.

4. From an applications perspective, an established link exists between ASPIC$^+$ and concrete, linguistic argumentation [Bex et al., 2010, 2013], meaning that, in principle, the work presented in this thesis can be applied to real argument data.

However, the use of ASPIC$^+$ is also in many ways for illustrative purposes only; the intention of the thesis is to show that the dynamics of existing systems of argumentation can be examined through a separate, decoupled model without needing to redefine or extend the system to accommodate the model.

1.4 Thesis structure

The thesis is structured as follows: Chapter 2 consists of a review of related literature, intended to place the work in context with respect to argumentation, belief revision and multi-agent dialogue, and to survey the existing connections between them. Chapter 3

contains formal definitions for the system of structured argumentation that is used in the thesis, as well as specifying new meta-level extensions to the system. A dialogue framework for a simple persuasion dialogue is specified in Chapter 4, which is subsequently used to illustrate the application of the model for *Argument Revision*. This chapter also introduces a running example, which is picked up again later in the thesis. The core of this work is the model for *Argument Revision*, which is specified in Chapter 5. Chapter 6 returns to and continues the running example by illustrating the use of *Argument Revision* in the context of dialogue. Finally, Chapter 7 concludes the thesis and identifies potential areas for future work.

Chapter 2

Background

2.1 Introduction

This chapter provides a review of related literature and is designed to place the thesis in context, in terms of argumentation, belief revision, multi-agent dialogue and the connections that exist between them.

The chapter is structured such that the three areas of argumentation, belief revision and dialogue will first be examined individually, with the final section exploring existing connections between them.

2.2 Argumentation

Argumentation is a multi-disciplinary field, attracting contributions from computer scientists, philosophers (e.g. [Walton, 1998, Reed and Tindale, 2010], discourse analysts and legal experts (e.g. [Bench-Capon and Prakken, 2010, Bex, 2011]). Work on argumentation ranges from the exploration of linguistic models of argument, including argument analysis using tools such as such as Araucaria [Reed and Rowe, 2004], OVA [Snaith et al., 2010] or Rationale [van Gelder, 2007], to the investigation of abstract, mathematical models, such as Dung [1995] abstract frameworks. However, recent work has seen the distinction between natural and abstract argumentation blurred, with the Argument Interchange Format (AIF)

[Chesñevar et al., 2006, Reed et al., 2008, Rahwan and Reed, 2009] serving as a bridge between concrete, linguistic analyses and abstract, computational models.

Computational models of argument have found a rapid increase in popularity in recent years, thanks in part to the successful COMMA series of conferences [Dunne and Bench-Capon, 2006, Besnard et al., 2008, Baroni et al., 2010, Verheij et al., 2012]. In this section, an emphasis will be placed on computational models of argument, specifically the work of Dung [1995] and subsequent extensions and applications to this theory.

2.2.1 Argumentation Frameworks

The work of Dung [1995] has proved highly influential in the field of argumentation. In his paper, Dung provides an abstract account of argumentation, consisting of two sets: a set of arguments, and a set of attack relations between them. No consideration is given to internal structure, nor the nature of attacks between arguments.

At the core of the theory is an argumentation framework.

Definition 2.2.1 ***Argumentation Framework***

An argumentation framework is a pair $AF = \langle AR, Atts \rangle$*, where:*

- AR *is a set of arguments*
- $Atts \subseteq AR \times AR$*, a set of attacks*

If for some pair of arguments $A, B \in AR$, $(A, B) \in Atts$, we say that A *attacks* B, which we shall subsequently denote as $Atts(A, B)$. A visual representation of an argumentation framework is shown through Example 2.2.1.

Example 2.2.1 *Consider an argumentation framework* $AF = \langle AR, Atts \rangle$ *, with:*

- $AR = \{a, b, c, d, e\}$
- $Atts = \{(a, a), (b, c), (c, b), (d, c), (e, d)\}$

This is rendered visually in Figure 2.1

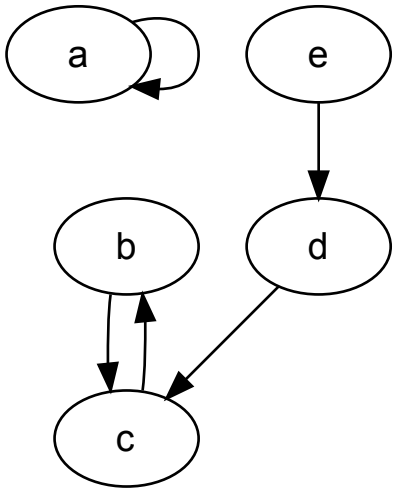

Figure 2.1: Framework for Example 2.2.1

Argument acceptability

The usefulness of argumentation frameworks comes from the notion of argument acceptability. Acceptability of arguments in frameworks is a cornerstone of Dung's work and, broadly speaking, describes a concept of consistency — given an argumentation framework, what arguments are consistent with each other, with respect to the attack relations between them?

A variety of different acceptability semantics exist (and will be explored later in this section), but all depend on the concept of a conflict-free set of arguments.

Definition 2.2.2 ***Conflict-free sets***

A set S of arguments if conflict-free iff $\nexists A, B \in S$ such that $(A, B) \in Atts$.

In example 2.2.1, it can be seen that the sets $\emptyset, \{b\}, \{c\}, \{d\}, \{e\}, \{b, d\}, \{c, e\}$ and $\{b, e\}$ are all conflict-free. However, this in itself is not very useful, because unless an argument attacks itself (as in the case of a in the example), it will always appear in at least one, possibly singleton, conflict-free set. A definition of *acceptability* is therefore used to

refine the idea of conflict-freeness.

Definition 2.2.3 *(Argument acceptability)*
*An argument $A \in AR$ is acceptable with respect to a set of arguments S iff for each argument $B \in AR$, if B attacks A then B is attacked by S. A conflict-free set of arguments is **admissible** iff each argument in S is acceptable with respect to S.*

Remark 1 *Some authors, such as Caminada [2007a], characterise acceptability as a function, $F(Args) = \{A \mid A$ is defended by $Args\}$*

In example 2.2.1, the sets $\emptyset, \{b\}, \{c\}, \{e\}, \{c, e\}$ and $\{b, e\}$ are all admissible. Two conflict-free sets, $\{c\}$ and $\{d\}$ are not admissible because c is attacked by d and there is no argument $\phi \in \{c\}$ such that ϕ attacks d; similarly, d is attacked by e and there is no argument $\phi \in \{d\}$ such that ϕ attacks e.

Having identified those sets that are admissible, the next step in argument evaluation is determining what arguments are acceptable with respect to the framework as a whole. This is achieved through various sceptical and credulous *acceptability semantics*. When a framework is evaluated under a certain semantics, one or more admissible sets of arguments are yielded, which conform to the properties of the chosen semantics. Each set of arguments is known as an *extension*.

All semantics are ultimately based on complete semantics. A set of arguments S is a complete extension if all arguments that are acceptable with respect to S belong to S.

Definition 2.2.4 *(Complete semantics) An admissible set S of arguments is called a complete extension iff each argument which is acceptable with respect to S belongs to S.*

Remark 2 *Using the function-based notation of Caminada [2007a], if $F(S) = S$, S is a complete extension (i.e. it is a fixed point).*

In example 2.2.1, $\{e\}$, $\{c, e\}$ and $\{b, e\}$ are complete extensions; $\emptyset$, $\{b\}$ and $\{c\}$ are not complete extensions (despite being admissible sets) because $\{e\}$ is not attacked by

any argument, and is therefore defended by any set of arguments. This means it must be a member of every complete extension.

Complete semantics on their own are of limited use, because they yield multiple extensions with no consideration for maximality. As can be seen in example 2.2.1, $\{e\}$ is a complete extension, but this ignores that when e is acceptable, either b or c (but not both) is also acceptable. Furthermore, it is also useful in certain applications (such as when abstract frameworks are used as a basis for beliefs) to have only one extension. To address these shortcomings, complete semantics are further refined into other semantics; while Dung [1995] provides formal definitions of these, it is convenient to summarise them as in definition 2.2.5.

Definition 2.2.5 *(Acceptability semantics)*
A conflict-free set of arguments S is:

- ***Admissible** if $S \subseteq F(S)$*
- *A **complete extension** if $S = F(S)$*
- *A **preferred extension** if S is a maximal (with respect to set inclusion) complete extension*
- *The **grounded extension** if S is the minimal (with respect to set inclusion) complete extension*
- *A **stable extension** if S attacks every argument in $AR \backslash S$*

In example 2.2.1, the preferred extensions are $\{b, e\}$ and $\{c, e\}$ and the grounded extension is $\{e\}$. Note that there is no stable extension, because for a set of arguments S to be a stable extension, it must be conflict-free and defeat every argument not in the extension. However, because argument a is self-defeating, it cannot appear in any conflict-free set and thus no such set exists that defeats every argument that is not a member of it.

This shows a significant drawback of stable semantics, in that they do not take relevance into consideration — a interacts with no arguments other than itself, but prevents any stable

extension from existing. Caminada [2006] proposes a solution to this problem through defining *semi-stable* semantics.

Definition 2.2.6 *(Semi-stable semantics)*
Let $AF = \langle AR, Atts \rangle$ be an argumentation framework and for some $S \subseteq AR$, S^+ be the set of all arguments defeated by S. S is a semi-stable extension if S is a complete extension and $S \cup S^+$ is maximal (with respect to set inclusion)

Returning again to example 2.2.1, the semi-stable extensions are $\{b, e\}$ and $\{c, e\}$.

In later work, Caminada [2007a] argues that further new acceptability semantics are needed to resolve the dilemma of certain semantics, such as preferred, yielding more than one extension. While one solution to this problem involves taking the intersection of the multiple extensions, in this approach the properties of the individual extensions are lost. Furthermore, in some semantics, such as preferred, taking the intersection (known as "sceptical-preferred") can result in a set which is not admissible. This was identified by Dung et al. [2007], and illustrated using the framework shown in Figure 2.2.

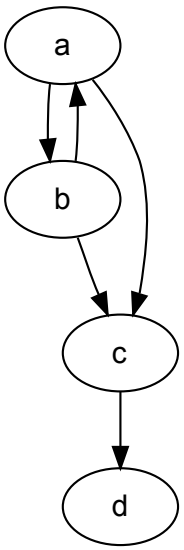

Figure 2.2: Framework illustrating the problem with sceptical-preferred semantics

There are two preferred extensions in this framework $\{a, d\}$ and $\{b, d\}$, with the sceptical-preferred extension being $\{d\}$. However, $\{d\}$ is not admissible, because d is attacked by c and there is no other argument in the set that defends d from c. Resolving this problem leads to *ideal semantics*:

Definition 2.2.7 *(Ideal semantics)*
The ideal extension is the maximal (w.r.t. set inclusion) admissible set that is a subset of each preferred extension

In Figure 2.2, the ideal extension is $\emptyset$, because $\{d\}$ is not admissible and there exists no other set that is a subset of both preferred extensions. It is proven by Dung et al. [2007] that the ideal extension, in common with all other extensions examined in this section, is a complete extension.

Similar to ideal semantics, Caminada [2007a] defines *eager semantics*, in terms of semi-stable semantics:

Definition 2.2.8 *(Eager semantics)*
The eager extension is the maximal (w.r.t. set inclusion) admissible set that is a subset of each semi-stable extension

Alternative approaches to solving the problem of multi-extension semantics have also been proposed. These will be examined in section 2.2.2.

2.2.2 Extensions, instantiations and implementations of Dung [1995]

Dung's [1995] work has proven highly significant in computational argumentation, illustrated by the considerable and varied body of work that has either extended it, or used it as a foundation. In this section, some of these extensions and instantiations will be explored.

Preferences

While Caminada [2007a] believes that the problem of multi-extension semantics can be resolved through defining new semantics, an alternative approach is to apply *preferences* to

arguments. In a preference-based approach, the success or otherwise of an attack depends on the attacking argument not being less-preferred to the attacked argument[1].

The first attempt at applying preferences to Dung-style frameworks was made by Amgoud and Cayrol [1997], where they develop a Preference Argumentation Framework (PAF) in which values are applied to arguments. Each argument holds a different value, with higher-valued arguments being more preferred to those with lower values. A similar approach is adopted by Bench-Capon and Dunne [2002] in their Value-based Argumentation Frameworks (VAFs), however a key difference between PAFs and VAFs is that the latter allows different arguments to hold the same value. A PAF can, therefore, be considered a special form of a VAF where each argument has a different value.

VAFs are defined as follows:

Definition 2.2.9 *(Value-based Argumentation Frameworks)*
A Value-Based Argumentation Framework (VAF) is a tuple
$VAF = \langle AR, Atts, V, val, valpref \rangle$ *where:*

- AR *and* $Atts$ *are the same as found in Dung's framework*
- V *is a non-empty set of values*
- val *is a mapping between* AR *and* V
- $valpref$ *is a preference relation expressing which values are preferred over others*

A criticism of value-based approaches is the question of how the values are arrived at. Modgil [2009a,b] provides an answer to this question through Extended Argumentation Frameworks, where instead of applying values to arguments, *preference arguments* are introduced that can attack the attacks between arguments.

[1]Note that this does not correspond to the attacking argument being more preferred — in some frameworks, the two arguments can hold equal preference and the attack succeeds, provided the attack is only in one direction.

Definition 2.2.10 *(Extended Argumentation Framework)*
An Extended Argumentation Framework is a tuple $EAF = \langle AR, Atts, D \rangle$ *where* AR *and* $Atts$ *are the same as found in Dung's framework and:*

- $D \subseteq AR \times Atts$
- *If* $(z,(x,y)),(z',(y,x)) \in D$ *then* $(z,z'),(z',z) \in Atts$

Remark 3 *The second condition states that two arguments which attack opposing attacks also attack each other.*

An advantage of Modgil's approach is that the preferences are themselves arguments (in AR), which allows preferences to be attacked in the same way as "normal" arguments.

Bipolarity

Amgoud et al. [2008] believe that a drawback of Dung's theory is its failure to capture the idea of support. This prompted them to specify *Bipolar Argumentation Frameworks* (BAFs) which introduce a new support relation between arguments.

Definition 2.2.11 *(Bipolar Argumentation Framework)*
A bipolar argumentation framework is a tuple $BAF = \langle AR, Atts, R \rangle$ *where* AR *and* $Atts$ *are the same as found in Dung's framework and* $R \subseteq AR \times AR$*, a support relation between arguments.*

Thus if $(a,b) \in R$, we say that a supports b.

The idea of support used in BAFs has come under criticism, as noted by Oren [2007]. However, he also argues in his thesis that despite these criticisms, support can play an important role when, for instance, considering evidence in an argumentation framework.

The ASPIC and ASPIC$^+$ frameworks

Instantiations of Dung's theory involve using it as the basis for less abstract frameworks. One such instantiation is the ASPIC$^+$ framework characterised by Prakken [2010] which

extended the original ASPIC framework of [Amgoud et al., 2006]. The ASPIC$^+$ framework incorporates structured arguments through combining the work of Vreeswijk [1997] with the work of Pollock [1987] on defeasible reasoning (see section 2.2.3).

The ASPIC$^+$ framework extends the original ASPIC framework in four ways:

1. The addition of a third way of attack — undermining
2. Abstracting the notion of contradiction between formulae to an abstract notion of contrariness, which is not necessarily symmetric (i.e. ϕ may be a contrary of ψ, but not vice versa)
3. The introduction of a knowledge base, with four types of premise
4. The introduction of preference orderings over the knowledge base and defeasible rules, which in turn leads to a preference ordering over arguments

The majority of the work presented in this thesis will be based on the ASPIC$^+$ framework, and as such the relevant formal definitions are provided in Chapter 3, along with other important logical preliminaries. However, informally, the fundamental basis of the framework is an *Argumentation System*, $\mathcal{AS} = \langle \mathcal{L}, ^- , \mathcal{R} \rangle$, where $\mathcal{L}$ is a logical language, $^-$ is a function from $\mathcal{L}$ to $2^{\mathcal{L}}$ representing contrariness between formulae and $\mathcal{R}$ is a set of strict ($\mathcal{R}_s$) and defeasible ($\mathcal{R}_d$) inference rules.

Prakken [2010] proves that, under certain assumptions, the ASPIC$^+$ framework satisfies Caminada's [2007a] rationality postulates:

- **Closure under subarguments:** for every argument in an extension, all its subarguments are also in the extension
- **Closure under strict rules:** the set of conclusions of all arguments in an extension is closed under strict-rule application
- **Direct consistency:** the set of conclusions of all arguments in an extension is consistent

- **Indirect consistency:** the closure of the set of conclusions of all arguments in an extension under strict-rule application is consistent

For these postulates to hold for ASPIC$^+$, Prakken assumes that an argumentation theory is *well-formed*. The formal definition of this is provided in Chapter 3, but informally an argumentation theory is well-formed iff (i) no formula is the consequent of an axiom; or (ii) no formula is the consequent of a strict rule.

A significant advantage of the ASPIC$^+$ framework is that, given a set of structured arguments, a Dung-style abstract framework can be derived then evaluated using the standard semantics. This allows the acceptability of structured arguments to be ascertained. A recent development by Bex et al. [2010, 2013] connected the Argument Interchange Format (AIF) [Chesñevar et al., 2006] to ASPIC$^+$, allowing real, analysed arguments expressed in the former to be translated into the logical concepts of the latter. This allows for the first time real, textual arguments to be analysed under Dung-style semantics in order to evaluate their acceptability.

The ASPIC$^+$ framework has received criticism from Amgoud [2012], who claims that it has five weaknesses, summarised as follows:

- The logical formalism is ill-defined
- The system may thus lead to undesirable results
- The system is grounded on several assumptions which may appear counter-intuitive
- The system may violate the rationality postulates on closure and direct consistency
- The system returns results which may not be closed under classical logic

Prakken and Modgil [2012] respond to these claims by providing counter-examples to the criticisms levied. The main principle of this response is focused on the nature of ASPIC$^+$ as a framework for specifying systems, and not a system itself. This means, Prakken and Modgil argue, that instantiating the framework in a way that violates the

rationality postulates (as Amgoud has done) demonstrates the aim of ASPIC$^+$ — to determine whether a certain instantiation satisfies or violates the postulates.

The ASPIC$^+$ framework has subsequently been extended and generalised by Modgil and Prakken [2013], however the work presented in this thesis will continue to use the framework as specified by Prakken [2010].

Other applications

In this section, other areas in which Dung's theory has been used are briefly examined.

Devereux and Reed [2009] and Devereux [2011] employ Dung-style argumentation frameworks as the basis for agent knowledge bases. In this work, the grounded extension was used as a simple means of deriving an agent's beliefs from an underlying framework, with beliefs defined simply in terms of arguments (e.g. if $\{a, b\}$ is the grounded extension of an argumentation framework, the agent believes $\{a, b\}$). This principle can be extended using the ASPIC$^+$ framework, such that instead of an agent believing the arguments in an extension, they believe the conclusions of those arguments, which will be a set of formulae in a language.

The abstract nature of Dung's work has seen it applied in areas where there exist some kind of entities (not necessarily arguments) and conflict between them. For instance, Bulling et al. [2009] use argumentation frameworks to model coalition structures in multi-agent systems, with the arguments representing agents, the attacks representing conflict between them and the acceptability semantics yielding possible coalitions (i.e. a set of agents (arguments) in a certain extension can form a coalition).

Implementations

The prominence of Dung's abstract frameworks has given rise to implementations of several systems based on the work. South et al. [2008] implemented *Dungine*, a Java-based reasoner that allows Dung-style frameworks to be constructed then evaluated under grounded, preferred and preferred-sceptical semantics. Similar to *Dungine* is OVA-gen [Snaith et al.,

2010], which provides an online interface for constructing abstract frameworks, which can then be evaluated under several different semantics, including grounded, preferred, ideal and semi-stable using algorithms based on those of [Vreeswijk, 2006, Caminada, 2007b, Dung et al., 2007].

Finally, TOAST [Snaith and Reed, 2012] is an online implementation of ASPIC$^+$, which shares a reasoning engine with OVA-gen. TOAST allows structured argumentation frameworks to be constructed using a simple propositional language and evaluated under a number of semantics. A connection between TOAST and the AIFdb [Lawrence et al., 2012b], using the work of [Bex et al., 2010, 2013], allows real arguments represented in the Argument Interchange Format to be processed using Dung-style acceptability semantics.

2.2.3 Defeasible argumentation

Defeasible reasoning was introduced by Pollock [1987] and divides rules of inference into two types: strict and defeasible. Informally, if a rule is strict, it holds *without exception*; if a rule is defeasible, it *usually* holds, but there may be some, possibly unknown, exceptional circumstances in which it does not.

A classic example used to illustrate the idea of defeasible reasoning is that of tweety the bird [Reiter, 1980].

Example 2.2.2 *Consider the following knowledge base, where $\Rightarrow$ represents a defeasible rule and $\rightarrow$ represents a strict rule:*

$$\mathcal{K} = \left\{ \begin{array}{l} bird(X) \Rightarrow flies(X), \\ penguin(X) \rightarrow bird(X), \\ penguin(X) \rightarrow \neg flies(X), \\ penguin(tweety) \end{array} \right\}$$

If we were to compute inferences over this knowledge base, we would arrive at two conclusions: that tweety, being a penguin, does not fly; but also that tweety, being a

penguin and hence a bird, does fly. However, while it is the case that, strictly, all penguins are birds and, strictly, all penguins do not fly, that "all birds fly" is only defeasible — that is, it is only in general that birds fly, but there are some exceptions (in this case, where the bird is a penguin).

Defeasible reasoning has found applications in computational models of argument. An insight into the use of defeasible reasoning in various systems of argumentation is provided by Prakken and Vreeswijk [2002]. One of their main observations is the way in which different systems handle defeasible consequence, which in some cases leads to unintuitive results.

More recently, García and Simari [2004] specify Defeasible Logic Programming (DeLP), a formalism that combines defeasible reasoning with logic programming. Given a set of information, DeLP performs a dialectical analysis to determine whether or not there is a warranted argument for a given query, q. An argument is warranted if it is not defeated by another argument in the dialectical analysis.

As noted in Section 2.2.2, the ASPIC$^+$ framework incorporates defeasible reasoning in providing structure to Dung [1995]-style argumentation frameworks. Similar to DeLP, rules in ASPIC$^+$ are categorised into strict and defeasible, with the latter being susceptible to *undercutting*, an attack by an argument that says the rule does not presently apply.

In both systems (DeLP and ASPIC$^+$), defeasibility is necessary for conflict to be resolved in all but the most trivial of examples. In DeLP, a strict argument for a certain conclusion will defeat a defeasible argument for a contradictory conclusion; in ASPIC$^+$, strict arguments are, in general, stronger than defeasible arguments and thus are preferred.

2.3 Belief revision

Belief revision is the study of how a knowledge base is updated when removing existing information, accommodating new information or a combination of both. One of the most influential theories of belief revision is the eponymous AGM theory of [Alchourrón et al.,

1985], where they propose the *AGM Postulates* for what they consider to be three types of change that can be made to a knowledge base:

Expansion: a new sentence Φ is added to a belief system K, together with its logical consequences. This is denoted as $K + \Phi$.

Revision: a new sentence Φ that is inconsistent with K is added, but further changes are made such that consistency is maintained. This is denoted as $K \dot{+} \Phi$

Contraction: a sentence Φ in K is retracted without adding any new sentences. To maintain closure under logical consequences, some other sentences may need to be given up. This is denoted as $K \dot{-} \Phi$

The Levi identity [Levi, 1977] allows revision to be expressed in terms of contraction and expansion, where $K \dot{+} \Phi = (K \dot{-} \neg\Phi) + \Phi$.

One of the main principles of belief revision is that of *minimal change* — when changing a belief set, it should be done with the fewest changes to the remaining or existing information. However, minimal change is not measured solely in terms of logical consequences [Gärdenfors, 1988, 1992]; instead, a qualitative *entrenchment ordering* is applied to beliefs, with those beliefs that possess the lowest degree of entrenchment being more willingly given up. As will be explored later in this section, however, this approach to minimality has a significant drawback.

2.3.1 AGM Postulates

The AGM Postulates, named after Alchourrón, Gärdenfors and Makinson [Alchourrón et al., 1985], govern the process of revising and contracting belief sets. Expansion does not require any postulates, because the process is self-evident — a belief set is expanded by adding the new sentence to it, with no consideration of the effects (e.g. causing inconsistency).

There are eight postulates each for revision and contraction, which are shown in tables 2.1 and 2.2 respectively, where $K_\perp$ is shorthand for "K is consistent" and $\vdash \Phi$ is shorthand for "Φ is logically possible".

$(K\dot{+}1)$	The outputs of the revision function are belief sets	$K\dot{+}\Phi$ is a belief set
$(K\dot{+}2)$	The input sentence Φ is accepted in $K\dot{+}\Phi$	$\Phi \in K\dot{+}\Phi$
$(K\dot{+}3)$ $(K\dot{+}4)$	Defined together and cover the case $\neg\Phi \notin K$	$K\dot{+}\Phi \subseteq K\dot{+}\Phi$ If$\neg\Phi \notin K$, then $K + \Phi \subseteq K\dot{+}\Phi$
$(K\dot{+}5)$	$K\dot{+}\Phi$ is consistent	$K\dot{+}\Phi = K_\perp$ iff $\vdash \neg\Phi$
$(K\dot{+}6)$	Revisions are based on the *knowledge* level and not the *analytic*	If $\vdash \Phi \leftrightarrow \Psi$, then $K\dot{+}\Phi = K\dot{+}\Psi$
$(K\dot{+}7)$ $(K\dot{+}8)$	The minimal change of K by Φ and Ψ is $K\dot{+}\Phi \wedge \Psi$	$K\dot{+}\Phi \wedge \Psi \subseteq (K\dot{+}\Phi) + \Psi$ If $\neg\Psi \notin K\dot{+}\Phi$, then $(K\dot{+}\Phi) + \Psi \subseteq K\dot{+}\Psi \wedge \Phi$

Table 2.1: AGM postulates for revision

$(K\dot{-}1)$	The outputs of the revision function are belief sets	$K\dot{-}\Phi$ is a belief set
$(K\dot{-}2)$	No new beliefs occur in $K\dot{-}\Phi$	$K\dot{-}\Phi \subseteq K$
$(K\dot{-}3)$	When $\Phi \notin K$, nothing is retracted	If $\Phi \notin K$, $K\dot{-}\Phi = K$
$(K\dot{-}4)$	$K\dot{-}\Phi \nvdash \Phi$	If not $\vdash \Phi$, then $\Phi \notin K\dot{-}\Phi$
$(K\dot{-}5)$	The *recovery postulate*; contractions can be 'undone'	If $\Phi \in K$, then $K \subseteq (K\dot{-}\Phi) + \Phi$
$(K\dot{-}6)$	Contractions are based on the *knowledge* level and not the *analytic*	If $\vdash \Phi \leftrightarrow \Psi$, then $K\dot{-}\Phi = K\dot{-}\Psi$
$(K\dot{-}7)$ $(K\dot{-}8)$	The minimal change of K by Φ and Ψ is $K\dot{-}\Phi \wedge \Psi$	$K\dot{-}\Phi \cap K\dot{-}\Psi \subseteq K\dot{-}\Phi \wedge \Psi$ If $\Phi \notin K\dot{-}\Phi \wedge \Psi$, then $K\dot{-}\Phi \wedge \Psi \subseteq K\dot{-}\Phi$

Table 2.2: AGM postulates for contraction

2.3.2 Epistemic entrenchment

When carrying out a belief revision process, it is difficult to uniquely specify a revision function, because each goal (i.e. revision or contraction) may have multiple ways through which it can be achieved [Gärdenfors, 1992]. One of the ways in which a method can be chosen is to examine what brings about the *minimal change* to a belief set.

In the AGM theory, minimal change is not measured solely in terms of logical consequences, but also employs a qualitative *entrenchment ordering* over beliefs — those beliefs with a lower degree of entrenchment are more willingly given up; in other words, giving them up is considered minimal.

Epistemic entrenchment is governed by five postulates [Gärdenfors, 1992]:

(EE1) If $\Phi \leq \Psi$ and $\Psi \leq \chi$, then $\Phi \leq \chi$ (transitivity)

(EE2) If $\Phi \vdash \Psi$, then $\Phi \leq \Psi$ (dominance)

(EE3) For any Φ and Ψ, $\Phi \leq \Phi \wedge \Psi$ or $\Psi \leq \Phi \wedge \Psi$ (conjunctiveness)

(EE4) When $K \neq K_{\perp}$, $\Phi \notin K$ iff $\Phi \leq \Psi$, for all Ψ (minimality)

(EE5) If $\Psi \leq \Phi$ for all Ψ, then $\vdash \Phi$ (maximality)

Table 2.3 provides the justifications that Gärdenfors [1992] gives for **(EE2)**-**(EE5)** (**(EE1)** is not provided with justification, because transitivity is obvious).

That the entrenchment ordering is qualitative is a significant drawback and represents a major issue in belief revision, because for an autonomous agent in a dynamic environment, it is not clear how new information can be accommodated in the entrenchment ordering. There may be some form of evaluation criteria built into the agent's reasoning mechanisms, but this assumes a closed-world principle, in which the agent is aware of every piece of information that may be presented to it. While this may be adequate in systems with very narrow applications, it is insufficient when considering, for instance, virtual participants on the proposed Argument Web [Rahwan et al., 2007] where, even if a debate is limited to a specific topic, in principle the exact information received by agents can vary considerably.

(EE2)	"...if Φ logically entails Ψ, and either Φ or Ψ must be rejected from K, then it will be a smaller change to give up Ψ and retain Φ rather than to give up Ψ, because Φ must be retracted too..."
(EE3)	"If one wants to retract $\Phi \wedge \Psi$ from K, this can only be achieved by giving up either Φ or Ψ and, consequently, the information loss incurred by giving up $\Phi \wedge \Psi$ will be the same as the loss incurred by giving up Φ or that incurred by giving up Ψ."
(EE4)	"...sentences already not in K have minimal epistemic entrenchment in relation to K..."
(EE5)	"...only logically valid sentences can be maximal in $\leq$."

Table 2.3: Justifications for **(EE2)**-**(EE5)**

2.4 Dialogue

A dialogue is a discussion of some kind between two or more participants[2]. The dialogue typology of Walton and Krabbe [1995] details six dialogue types, with each type having an initial situation, a main goal (the purpose of the dialogue) and individual goals (what each participant aims to achieve from the dialogue).

The initial situation in **persuasion dialogue** is one of conflict, where two or more participants take a contrary opinion. The main goal of this dialogue is to resolve the conflict, with the individual goal of each agent being to have their opinion accepted.

A **negotiation dialogue** starts with a need for co-operation, with a main goal of making a deal. The individual goal of each agent is to get the best possible deal for itself.

An **inquiry dialogue** arises out of general ignorance, and has as its main goal the growth of knowledge (consider real-life inquires, such as those carried out by the police or judiciary). Each agent's individual goal is to either find or destroy a proof which will ultimately influence the outcome of the inquiry.

A **deliberation dialogue** has a need for action as its initial situation and its main goal is to reach a decision on which action to take. The goal of each agent is to influence the outcome.

[2]While "dialogue" might suggest exactly two participants, in the context of philosophical and computational models of dialogue, this is extended to "two *or more*".

An **information-seeking dialogue** starts from a position of general ignorance, and aims to spread knowledge. Each agent's individual goal is to spread or hide knowledge.

The sixth and final dialogue in Walton & Krabbe's typology is "Eristics", which is analogous to physical fighting. In the context of multi-agent systems research, this dialogue can be overlooked.

2.4.1 Dishonesty in dialogue

Lying can be considered a fundamental human behaviour, however research into its use in artificial intelligence is somewhat limited. Sklar et al. [2005] provide a model for contradiction in dialogue, which allows a participant to support an argument with information they might not believe. The question of "why lie?" is answered with several scenarios; one is that offering an acceptable argument for a lie q that supports the statement p is (somehow) easier than finding an acceptable argument for p.

Sakama et al. [2010] look to provide a formal account of lying, by specifying what they see as three types of dishonesty:

- **Lying:** the process of uttering a believed-false statement with the intention that the hearer believes it, and believes that the speaker believes it
- **Bullshit:** (henceforth referred to as "BS") the process of uttering a statement that is neither believed to be true nor false, with the intention that the hearer believes it, and believes that the speaker believes it.
- **Deceit:** the process of uttering a statement that is believed-true, but with the intention that the hearer uses it to draw a false conclusion

One of the key elements of Sakama et al.'s specification of dishonesty is that dishonest acts must be kept as small as possible so as to reduce the risk the dishonesty being exposed. The size of dishonesty is measured in terms of the impact on other beliefs — lying (in the formal sense) is uttering the opposite of something you believe (e.g. uttering $\neg\phi$ when ϕ is

believed), so anything that is a consequence of that belief (e.g. if $\phi \vdash \psi$) or that belief is a consequence of (e.g. $\theta \vdash \phi$) will be unavailable to the participant later in the dialogue, unless they can use another dishonest act to cover it up.

2.4.2 Dialogue in multi-agent systems

Multi-agent systems (MASs) are software systems built from multiple autonomous agents, which each have their own goals, and the ability to control their behaviour in achieving them [Wooldridge, 2001]. Communication between agents in a MAS is vital if the overall goal of the system is to be reached, with computational accounts of philosophical dialogues having been identified as one such means of communication.

The work of Parsons and Jennings [1996], Reed [1998] and McBurney and Parsons [2002] laid the groundwork for specifying computational protocols for argumentative inter-agent communication. Parsons and Jennings [1996] specify a formal protocol for negotiation between agents looking to find ways to solve problems, while Reed [1998] provides a computational account of the Walton and Krabbe [1995] dialogue typology (with the exception of eristics, which models physical conflict) and McBurney and Parsons [2002] formalised the modelling of dialogue games for agent communication.

2.5 Argumentation, belief revision and dialogue

In this section, existing connections between argumentation, belief revision and dialogue are explored.

2.5.1 Argumentation and dialogue

Connections between argumentation and dialogue are long-established thanks to the close link between two forms of argument — argument-as-product, which is sets of premises and inferences leading to conclusions, and argument-as-process, which is two or more

participants engaging in an argumentative dialogue (which in turn can lead to an argument-as-product structure).

From a philosophical standpoint, Walton [1998] illustrates how argument is employed in many dialogue types, and can be used to contribute to the goals of the dialogue. As noted in section 2.4.2, many computational dialogue games based on systems of argumentation have been specified, based (at least in part) on the Walton and Krabbe [1995] typology. Work by Parsons and Jennings [1996], Reed [1998] and McBurney and Parsons [2002], laid the groundwork for specifying computational protocols for argumentative inter-agent communication.

Prakken [2005] specifies a framework for argumentative dialogue and shows that dialogues based on it implicitly build an argument structure, which can be used to evaluate certain properties of the dialogue. This returns to the link between argument-as-process and argument-as-product.

More recently, persuasion dialogues have been specified by Devereux and Reed [2009], Devereux [2011], based on Walton and Krabbe's PPD, and uses Dung [1995] style abstract argumentation frameworks as a source of agent beliefs, and content of locutions.Weide and Dignum [2011] specify a simple persuasion dialogue based on Prakken's [2005] framework, and use the ASPIC$^+$ framework as the basis of beliefs and the content of locutions.

While persuasion appears to be the most popular dialogue type in argumentation, since it models most closely the idea of "having an argument", Black and Hunter [2007, 2009] specify a computational protocol for an inquiry dialogue using argumentation. An inquiry dialogue aims to establish a truth or find new knowledge, instead of one party trying to persuade the others to accept their position.

2.5.2 Argumentation and Belief Revision

The aim of belief revision is to resolve inconsistencies that may arise when accommodating new information in a belief set. The idea of inconsistency in argumentation is not clearly

defined, although some authors such as Besnard and Hunter [2008] do consider it, because the argumentation machinery should deal with statements or arguments that are in conflict with one another.

However, this does not preclude connections between argumentation and belief revision. Such connections can trace their roots to the Truth Maintenance System of Doyle [1979], which allows for beliefs to be represented and consistency restored if required, but more recently the exploration of connections has found new momentum, with two approaches to combining the fields having emerged. One is the use of argumentation to assist with belief revision, such as the work of Krümpelmann et al. [2012] on using deductive argumentation to perform selective revision [Fermé and Hansson, 1999, Falappa et al., 2012], or the work of Falappa et al. [2002] that uses defeasible reasoning and explanation in non-prioritized belief revision (where input sentences are not necessarily accepted [Hansson, 1999]); the other is the use of belief revision concepts to model the dynamics of argumentation systems, such as in Argument Theory Change [Rotstein et al., 2008].

Falappa et al. [2009, 2011] further develop the connections between argumentation and belief revision by developing a conceptual view of the topic. This conceptual view is first defined by four steps of reasoning:

- *Receiving new information*: new information can be in many forms, e.g. a propositional fact or an argument,
- *Evaluating new information*: Upon receipt of new information, an agent will evaluate it in order to be convinced it is true. If the information is an observation, truth is usually assumed. However, if it is communicated by another agent, some justification will be required.
- *Changing beliefs*: If the agent chooses to accept the information, it will apply belief revision techniques to incorporate it into its knowledge base
- *Inference*: From its new epistemic state, the agent derives plausible beliefs that guide its behaviour

Referring to these as "steps" implies that each is a distinct process, following on from the previous one. However, it is possible there may be overlap: for instance, one could consider the evaluation of new information as being a series of tests, each of which must pass for the information to be evaluated as being acceptable to the agent. One of these tests could be a determination of the impact (with respect to minimal change) this information would have on the agent's current beliefs — the initial steps of actually accommodating it.

The overall conclusions from Falappa et al.'s work is that argumentation and belief revision should not be seen as competing alternatives, but instead seen as complimentary; combining the two allows for a richer modelling of the reasoning process.

Two applications of belief revision to argumentation will be now be briefly examined.

Change in abstract argumentation

Cayrol et al. [2010] specify change operators for Dung-style argumentation frameworks, in terms of the effects of adding an argument and its interactions (i.e. attacks) to the framework. The changes are defined in terms of the effect on the set of extensions (under unspecified semantics) in the framework and are summarised in Table 2.4.

A key principle in the specification of Cayrol et al.'s change operators is that their determination is based on existing, quantifiable effects on the system of argumentation, in this case Dung-style abstract frameworks. They do not depend on modifications to the system that are specific to modelling and assessing change.

Argument Theory Change

Argument Theory Change (ATC) is an argumentation framework whose dynamics are captured through the application of belief change operators [Rotstein et al., 2008, Moguillansky et al., 2010]. ATC is an extension to Dung's (1995) theory by incorporating *subarguments* and a set of *active arguments* (those arguments that are available to perform reasoning). So if an argument is activated (i.e. becomes available for consideration), the set of active arguments is further modified to make this argument warranted.

Property for a change operation	Characterisation of the property
the change is **decisive**	$E = \emptyset$ or $E = \{\emptyset\}$ or $\mid E \mid > 2$ and $\mid E' \mid = 1$ and $E' \neq \{\emptyset\}$
the change is **restrictive**	$\mid E \mid > \mid E' \mid > 2$
the change is **questioning**	$\mid E \mid < \mid E' \mid$
the change is **destructive**	$E \neq \emptyset$ and $E \neq \{\emptyset\}$ $E' = \emptyset$ or $E' = \{\emptyset\}$
the change is **expansive**	$\mid E \mid = \mid E' \mid$ and $\forall \varepsilon_i^j \in E', \exists \varepsilon_i \in E, \varepsilon_i \subset \varepsilon_i^j$
the change is **conservative**	$E = E'$
the change is **altering**	$\mid E \mid = \mid E' \mid$ and $\exists \varepsilon_i \in E$ s.t. $\forall \varepsilon_j^i \in E', \varepsilon_i \not\subseteq \varepsilon_j^i$

Table 2.4: Structural properties for a change operation, from [Cayrol et al., 2010]

The basis for Argument Theory Change is a *Dynamic Argumentation Framework*:

Definition 2.5.1 *A dynamic abstract argumentation framework (DAF) is a tuple* $\langle \mathbb{U}, \hookrightarrow, \sqsubseteq \rangle$ $[\mathbb{A}]$*, where* $\mathbb{U}$ *is a finite set of arguments called* ***universal****,* $\mathbb{A} \subseteq \mathbb{U}$ *is called the* ***set of active arguments****,* $\hookrightarrow \subseteq \mathbb{U} \times \mathbb{U}$ *denotes the* ***attack relation*** *and* $\sqsubseteq \subseteq \mathbb{U} \times \mathbb{U}$ *denotes the* ***subargument relation***

Remark 4 $\mathbb{U}$*,* $\hookrightarrow$ *and* $\sqsubseteq$ *are considered static, while* $\mathbb{A}$ *is dynamic.*

ATC can be broadly described as the process of adding an argument to a theory, then applying belief revision techniques to make that argument warranted. Different approaches to bringing about the warrant are available, such as through activating defeaters [Moguillansky et al., 2010].

A limitation of ATC is that a model of dynamics is built in to the framework, and as such it is defined over a set of universal arguments, which is every argument that could be made available for reasoning in the system. This is sufficient for specific applications with limited domains where all possible arguments are known (for instance, reasoning about a certain medical condition), however in dialogical applications, while restricting the dialogue to a certain domain is not altogether unreasonable, a shared knowledge between the participants of all possible arguments limits the dialogue somewhat; for instance, it

would be difficult for a participant to implement certain strategies if their opponents were already aware of every possible argument.

2.5.3 Belief revision and dialogue

Despite what appear to be obvious connections, the overlap between belief revision and dialogue remains under-researched. Zhang et al. [2004] consider a negotiation dialogue as a form of mutual belief revision, where each participant examines and evaluates the others' beliefs as part of the process of reaching the goal. A similar technique is also used by Pilotti et al. [2012], where belief revision is used for the generation, selection and evaluation of arguments in a negotiation dialogue.

In certain protocols based on the Walton and Krabbe [1995] typology, each participant has a commitment store, to which utterances and concessions are added, and from which retractions are removed. As Girle [1997, 2002] identifies, a commitment store is like a belief set and as such the processes of incurring and retracting commitment can be seen as analogues for, respectively, expansion (or revision) and contraction.

In the dialogue game PPD_0 [Walton and Krabbe, 1995], when a participant retracts a statement from their commitment store, they must also retract any other statements of which the retracted is a consequence, in a process known as a *stability adjustment*. However, what is not made clear is that, when a choice of addition retractions exist (e.g. if $P, Q \vdash R$ and R is retracted, which out of P or Q is also retracted), what criteria should be used to choose between them? There is, again, a natural link back to belief revision — perform the retraction that brings about the minimal change.

2.6 Summary and discussion

This chapter has provided a review of relevant literature connected to argumentation, belief revision and dialogue.

In the case of argumentation, specific focus was made on computational models,

specifically Dung's [1995] abstract frameworks, along with their subsequent instantiations, extensions and applications. Belief revision, meanwhile, focused on the AGM model [Alchourrón et al., 1985], its concept of minimal change and why the traditional method of determining it is insufficient.

When looking at existing connections between argumentation and belief revision, two broad approaches are evident — the use of argumentation in assisting the belief revision process (such as the use of deductive argumentation to select a revision [Krümpelmann et al., 2012]) and the application of belief revision principles to systems of argumentation to model their dynamics (such as Argument Theory Change [Rotstein et al., 2008, Moguillansky et al., 2010]).

While the modelling of argument dynamics through belief change operators appears to have a solid foundation, what appears to be under-researched is a decoupling of the model for argument change from the framework to which it is applied. For instance, Argument Theory Change defines a new argumentation framework ("Dynamic Argumentation Framework"), with the revision operations being a part of it. Defining a new framework that incorporates its model of dynamics does have advantages; for instance, it allows the framework to be specified with the need to model dynamics in mind, which in turn can give a more specific determination of concepts such as minimal change. However, being able to model the dynamics of existing frameworks, without modifying them to incorporate the model, also has its benefits; from a theoretical perspective, it allows existing frameworks with established and proven properties to have their dynamics modelled without needing to modify the framework itself, which carries the risk of changing its properties. In terms of applications, decoupling the model of dynamics turns it into an additional, optional tool for processing the framework. This allows, for instance, a dialogue to be based on the framework, with some participants examining dynamics and others not.

What is perhaps most striking is the lack of research into the connections between dialogue and belief revision, despite the obvious overlaps between the dynamics of commitment stores [Walton and Krabbe, 1995] and the dynamics of belief sets [Gärdenfors,

1988, 1992], a link that appears to have been noted only by Girle [1997, 2002], but without any further significant action.

Commitment stores are, by Walton and Krabbe's definition, dynamic — as a dialogue progresses, each participant's store is updated, either through adding new statements, or removing previous ones, with the latter (in certain protocols) also incorporating a stability adjustment, where statements that allow the retracted statement to be concluded are also given up. The latter situation in particular represents an area well-suited to the application of belief revision techniques. While Walton and Krabbe describe what a stability adjustment is, and its intention, no consideration is given to *how* it should be performed, and if multiple possibly adjustments exist, which should be chosen. Considering this problem in the context of belief revision, it can be framed thus: when faced with the need to retract, what is the minimal set of retractions that i) removes the required statement and ii) prevents that statement from being inferred in the new commitment store?

It is also possible that a participant in a dialogue does not want to concede a statement to their opponent. If a concession requires a contraction and subsequent stability adjustment, it may be that any possible adjustment has a significant effect on remaining commitments, and potential future statements in the dialogue. So other options beyond concession and retraction must be considered, to give the participant a choice.

One such option is to be dishonest. Dishonesty is an important, but under-researched topic in artificial intelligence [Sakama et al., 2010], because it represents a fundamental human behaviour. Sakama et al.'s logical account of lying places importance on "minimal lies" — tell only those lies that have a minimal impact on other beliefs, because those are the easiest to maintain. This concept of minimality also has obvious links to belief revision; if dishonesty were considered a form of belief revision (and so a revision process is performed with respect to the source of the dishonesty), then minimal change can be used as a means of determining minimal lies.

In Chapter 5 a model for argument revision is specified which addresses the issues identified in this review. The model is decoupled from the framework for argumentation

whose dynamics it captures, removing the need to respecify the framework and, subsequently, ensure its properties are maintained. In terms of dialogue, the model provides a participant with the ability to assess the impact of a concession then subsequent retraction, or dishonest act to continue defending their position.

Chapter 3

Logical preliminaries

3.1 Introduction

The purpose of this chapter is to introduce two important foundations that the remainder of the thesis shall build upon and use. The first is the ASPIC$^+$ framework [Prakken, 2010], an abstract framework for argumentation that instantiates the work of Dung [1995] by giving arguments structure. The second foundation is the provision of an extension to ASPIC$^+$ to incorporate meta-argumentation.

3.2 The ASPIC$^+$ framework

Instantiations of Dung's 1995 abstract theory of argumentation involve using it as the basis for less abstract frameworks. One such instantiation is the ASPIC$^+$ framework characterised by Prakken [2010], which also extended the work of the ASPIC project [Amgoud et al., 2006] and the work of Caminada and Amgoud [2007]. The ASPIC$^+$ framework incorporates structured arguments, through combining the work of Vreeswijk [1997] with the work of Pollock [1987] on defeasible reasoning.

The ASPIC$^+$ framework extended the original ASPIC framework in four ways:

1. The addition of a third way of attack — undermining

2. Abstracting the notion of contradiction between formulae to an abstract notion of contrariness between formulae, which is not necessarily symmetric (i.e. ϕ may be a contrary of ψ, and yet not the reverse)

3. The introduction of a knowledge base, with four types of premise

4. The introduction of preferences between arguments, defeasible inference rules (see [Pollock, 1987]) and the knowledge base.

The fundamental notion of the ASPIC$^+$ framework is an *argumentation system*:

Definition 3.2.1 *An argumentation system is a tuple $\mathcal{AS} = \langle \mathcal{L}, ^- , \mathcal{R}, \leq \rangle$ where:*

- *$\mathcal{L}$ is a logical language*
- *$^-$ is a contrariness function from $\mathcal{L}$ to $2^{\mathcal{L}}$*
- *$\mathcal{R} = \mathcal{R}_s \cup \mathcal{R}_d$ is a set of strict ($\mathcal{R}_s$) and defeasible ($\mathcal{R}_d$) inference rules such that $\mathcal{R}_s \cap \mathcal{R}_d = \emptyset$*
- *$\leq$ is a partial pre-order on $\mathcal{R}_d$*

Remark 5 *Given two formulae $\phi, \psi \in \mathcal{L}$, the notation $\phi \in \overline{\psi}$ means "ϕ is a contrary of ψ and the notation $\phi = -\psi$ means "ϕ and ψ are contradictory" (i.e. $\phi \in \overline{\psi}$ and $\psi \in \overline{\phi}$).*

A defeasible inference rule is a rule that *usually* holds, however there may be some exceptional circumstances where it does not; a strict inference rule is a rule that *always* holds, without exception.

An argumentation system contains a knowledge base, whose elements are formulae of the language $\mathcal{L}$:

Definition 3.2.2 *A knowledge base in an argumentation system $\langle \mathcal{L}, ^- , \mathcal{R}, \leq \rangle$ is a pair $\langle \mathcal{K}, \leq' \rangle$ where:*

- $\mathcal{K} \subseteq \mathcal{L}$

- $\mathcal{K} = \mathcal{K}_n \cup \mathcal{K}_p \cup \mathcal{K}_a$, *with:*
  - $\mathcal{K}_n$ *being a set of (necessary) axioms*
  - $\mathcal{K}_p$ *being a set of (ordinary) premises*
  - $\mathcal{K}_a$ *being a set of assumptions*
- $\leq'$ *is a partial pre-order on* $\mathcal{K} \backslash \mathcal{K}_n$

From the knowledge base ($\mathcal{K}$) and rules ($\mathcal{R}$), arguments are constructed. For an argument $\mathcal{A}$, $Prem(\mathcal{A})$ is a function that returns all the premises in $\mathcal{A}$; $Conc(\mathcal{A})$ is a function that returns the conclusion of $\mathcal{A}$; $Sub(\mathcal{A})$ is a function that returns all the sub-arguments of $\mathcal{A}$; $DefRules(\mathcal{A})$ is a function that returns all defeasible rules in $\mathcal{A}$; and $TopRule(\mathcal{A})$ is a function that returns the last inference rule use in $\mathcal{A}$.

$\mathcal{A}$ can be one of three types, and these functions in each case are:

1. When $\mathcal{A} = \varphi$, for $\varphi \in \mathcal{K}$, with:
   - $Prem(\mathcal{A}) = \{\varphi\}$
   - $Conc(\mathcal{A}) = \varphi$
   - $Sub(\mathcal{A}) = \varphi$
   - $DefRules(\mathcal{A}) = \emptyset$
   - $TopRule(\mathcal{A}) =$ undefined.
2. When $\mathcal{A} = \mathcal{A}_1, \ldots, \mathcal{A}_n \rightarrow \psi$ $\mathcal{A}_1, \ldots, \mathcal{A}_n$ such that there exists a strict rule $Conc(\mathcal{A}_1), \ldots Conc(\mathcal{A}_n) \rightarrow \psi$ in $\mathcal{R}_s$ with:
   - $Prem(\mathcal{A}) = Prem(\mathcal{A}_1) \cup \ldots \cup Prem(\mathcal{A}_2)$
   - $Conc(\mathcal{A}) = \psi$
   - $Sub(\mathcal{A}) = Sub(\mathcal{A}_1) \cup \ldots \cup Sub(\mathcal{A}_n) \cup \{\mathcal{A}\}$
   - $DefRules(\mathcal{A}) = DefRules(\mathcal{A}_1) \cup \ldots \cup DefRules(\mathcal{A}_n)$

- $TopRule(\mathcal{A}) = Conc(\mathcal{A}_1), \ldots, Conc(\mathcal{A}_n) \rightarrow \psi$.

3. When $\mathcal{A} = \mathcal{A}_1, \ldots, \mathcal{A}_n \Rightarrow \psi$ for arguments $\mathcal{A}_1, \ldots, \mathcal{A}_n$ such that there exists a defeasible rule $Conc(\mathcal{A}_1), \ldots Conc(\mathcal{A}_n) \Rightarrow \psi$ in $\mathcal{R}_d$ with

   - $Prem(\mathcal{A}) = Prem(\mathcal{A}_1) \cup \ldots \cup Prem(\mathcal{A}_2)$
   - $Conc(\mathcal{A}) = \psi$
   - $Sub(\mathcal{A}) = Sub(\mathcal{A}_1) \cup \ldots \cup Sub(\mathcal{A}_n) \cup \{\mathcal{A}\}$
   - $DefRules(\mathcal{A}) = DefRules(\mathcal{A}_1) \cup \ldots \cup DefRules(\mathcal{A}_n)$
   - $TopRule(\mathcal{A}) = Conc(\mathcal{A}_1), \ldots, Conc(\mathcal{A}_n) \Rightarrow \psi$.

Arguments are classified based on their rules and premises:

Definition 3.2.3 *An argument A is:*

- ***strict*** *if* $DefRules(A) = \emptyset$
- ***defeasible*** *if* $DefRules(A) \neq \emptyset$
- ***firm*** *if* $Prem(A) \subseteq K_n$
- ***plausible*** *if* $Prem(A) \not\subseteq K_n$

That is, an argument is strict if it has no defeasible rules, or defeasible otherwise, and an argument is firm if its premises consists only of axioms, or plausible otherwise.

An argument can be attacked in three ways: on a (non-axiom) premise (undermine), on a defeasible inference rule (undercut) or on a conclusion (rebuttal).

Given an argumentation system $\mathcal{AS}$ and a knowledge base $\mathcal{KB}$, an argumentation theory is $\mathcal{AT} = \langle \mathcal{AS}, \mathcal{KB}, \preceq \rangle$ where $\preceq$ is an argument ordering on the set of all arguments that can be constructed from $\mathcal{KB}$ in $\mathcal{AS}$. An argumentation theory is *well-formed* iff: if ϕ is a contrary of ψ then $\psi \notin \mathcal{K}_n$ and ψ is not the consequent of a strict rule [Modgil and Prakken, 2011].

Preferences over arguments are used to determine whether or not an attack succeeds (i.e. results in a defeat). Preference orderings are based on an idea of being "reasonable", which in turn is based on *maximal fallible sub-arguments*:

Definition 3.2.4 *(Maximal fallible sub-arguments)*
*For any argument $\mathcal{A}$, an argument $\mathcal{A}' \in Sub(\mathcal{A})$ is a **maximal fallible sub-argument** of $\mathcal{A}$ if:*

1. *$\mathcal{A}'$'s final inference is defeasible, or $\mathcal{A}'$ is a non-axiom premise; and*
2. *there is no $\mathcal{A}'' \in Sub(\mathcal{A})$ such that $\mathcal{A}'' \neq \mathcal{A}$, $\mathcal{A}' \in Sub(\mathcal{A})$ and $\mathcal{A}''$ satisfies condition (1)*

Remark 6 *For an argument $\mathcal{A}$, $M(\mathcal{A})$ is its maximal fallible sub-arguments.*

Definition 3.2.5 *(Reasonable argument orderings)*
*Let $\mathcal{A}$ and $\mathcal{B}$ be arguments with contradictory conclusions such that $\mathcal{B} \prec \mathcal{A}$. If the argument ordering $\preceq$ is **reasonable** then there exists a $\mathcal{B}_i \in M(\mathcal{B})$ and an $\mathcal{A}^+$ with $\mathcal{A} \in Sub(\mathcal{A}^+)$ such that $Conc(\mathcal{A}^+) = -Conc(\mathcal{B}_i)$ and $\mathcal{A}^+ \not\prec \mathcal{B}_i$*

Finally, ASPIC$^+$ defines several closure operators, two of which will be used in this thesis. Firstly, closure under transposition:

Definition 3.2.6 *(Transposition)*
A strict rule s is a transposition of $\varphi_1, \ldots, \varphi_n \rightarrow \psi$ iff $s = \varphi_1, \ldots, \varphi_{i-1}, -\psi, \varphi_{i+1}, \ldots, \varphi_n \rightarrow -\varphi_i$ for some $1 \leq \imath \leq n$.

Definition 3.2.7 *(Closure under transposition)*
Let $\mathcal{R}_s$ be a set of strict rules. $\mathcal{C}l_{tp}(\mathcal{R}_s)$ is the smallest set such that:

- *$\mathcal{R}_s \subseteq \mathcal{C}l_{tp}(\mathcal{R}_s)$, and*
- *If $s \in \mathcal{C}l_{tp}(\mathcal{R}_s)$ and t is a transposition of s, then $t \in \mathcal{C}l_{tp}(\mathcal{R}_s)$.*

If $\mathcal{C}l_{tp}(\mathcal{R}_s) = \mathcal{R}_s$, $\mathcal{R}_s$ is closed under transposition.

Secondly, the closure of a set of formulae under strict rule application:

Definition 3.2.8 *(Closure of a set of formulae)*
Let $P \subseteq \mathcal{L}$. The closure of P under the set $\mathcal{R}_s$ of strict rules, denoted $\mathcal{C}l_{R_s}(P)$, is the smallest set such that:

- $P \subseteq \mathcal{C}l_{R_s}(P)$
- *if* $\varphi_1, \ldots, \varphi_n \rightarrow \psi \in \mathcal{R}_s$ *and* $\varphi_1, \ldots, \varphi_n \in \mathcal{C}l_{\mathcal{R}_s}(P)$ *then* $\psi \in \mathcal{C}l_{R_s}(P)$

3.3 Meta-argumentation

In meta-argumentation, statements and reasoning can be made about arguments themselves. Modgil [2009a] employs meta-level argumentation as a means of reasoning about preferences in argumentation frameworks, which Modgil and Prakken [2010] combined with [Prakken, 2010] to allow for reasoning about preferences in structured argumentation frameworks. In all this work, preferences between arguments themselves become meta-level arguments.

Weide and Dignum [2011] also use meta-argumentation with the ASPIC$^+$ framework, only their model is not limited only to preferences. In their work, rules and arguments themselves can be elevated as meta-level arguments. However, their definition of a meta-argumentation system does not require that arguments about object-level concepts exist in the associated meta-level system; the only constraint is that the meta-language subsumes the object-level language, and also contains every object-level rule and argument as formulae. Preferences (between rules and formulae) and conflict are not part of the meta-language and thus arguments about them cannot be constructed. Furthermore, a binary predicate, $\preceq$, is explicitly added to the meta-language as a means of reasoning about conclusive force.

This approach to meta-argumentation has several drawbacks:

1. There is no requirement that object-level rules and arguments have corresponding arguments in a meta-argumentation system; all that is specified is that rules and arguments are a part of the meta-language. This limits the ability of meta-argumentation to be used to reason about the object-level system, because there is no guarantee that a given component at the object-level is represented by a meta-argument.

2. The addition of the binary predicate $\preceq$ to the meta-language is redundant, because preferences between arguments are already expressed at the object-level. Instead of introducing a new predicate, preferences can be represented at the meta-level to convey the same meaning.

In spite of these drawbacks, the general principle of representing object-level components (rules, preferences, contrariness and arguments) as formulae of a meta-language provides several advantages. Such advantages include the ability to reason about, and provide justification for, object-level components, and representing existing concepts such as undercutting in ways that do not place assumptions on the object language.

Consider as an example the following information about whether or not a person who is ill should visit their doctor.

> *If you are ill, you should visit your doctor. However, if you are ill and have the flu, it does not follow that you should visit your doctor. This is because flu is contagious and you could infect others.*

If we were to encode this in ASPIC$^+$, we can easily represent the first sentence as a rule: $[r1]\ ill \Rightarrow go\_to\_doctor$. The second sentence represents an exception to the rule (Chapter 2, Section 2.2.3), which Prakken [2010] represents as rules, by first assuming rule labels can be represented in the object language $\mathcal{L}$: $[r2]\ have\_flu \Rightarrow \neg[r1]$.

However, this exception has justification ("flu is contagious…"), which has no way of being represented. This is the role played by meta-argumentation — the exception becomes a formula in a meta-language which then allows it to be reasoned about in the

same way as formulae in the object language. Furthermore, as will be shown in Definitions 3.3.2 and 3.3.3, meta-argumentation also allows for a representation of exceptions that do not depend on rule labels being represented in the object language.

In expanding ASPIC$^+$ to account for meta-arguments, we address the shortcomings of Weide and Dignum [2011]'s approach by defining a meta-argumentation system as not being derived from an object-level argumentation system, but instead through a series of bijections, where components of the object-level system must be represented by a (possibly atomic) argument in the meta-level system, and arguments in the meta-level system result in components of the object-level system.

It should be stressed at this point that the purpose of this thesis is not to contribute significantly to the study of meta-argumentation, but will instead only be using it as a tool in dialogue and argument revision. Thus the process through which an object-level component becomes a meta-level argument, and vice versa, will be left implicit.

Definition 3.3.1 *(Meta-argumentation)*
Given an object-level argumentation system $\mathcal{AS}_i = \langle \mathcal{L}_i, \mathcal{R}_i, cf_i, \leq_i \rangle$ *and a meta-argumentation system* $\mathcal{AS}_{i+1} = \langle \mathcal{L}_{i+1}, \mathcal{R}_{i+1}, cf_{i+1}, \leq_{i+1} \rangle$*:*

- $\mathcal{L}_i \subseteq \mathcal{L}_{i+1}$
- $\mathcal{A} \in \mathcal{A}rgs(\mathcal{AS}_i) \leftrightarrow \lceil \mathcal{A} \rceil \in \mathcal{L}_{i+1}$
- $r \in \mathcal{R}_i \leftrightarrow \exists \mathcal{A} \in \mathcal{A}rgs(\mathcal{AS}_{i+1})$ *s.t.* $Conc(\mathcal{A}) = \lceil r \rceil$
- $\phi \in \overline{\psi}$ *in* $\mathcal{AS}_i \leftrightarrow \exists \mathcal{A} \in \mathcal{A}rgs(\mathcal{AS}_{i+1})$ *s.t.* $Conc(\mathcal{A}) = \lceil \phi \in \overline{\psi} \rceil$
- $\phi <' \psi$ *in* $\mathcal{AS}_i \leftrightarrow \exists \mathcal{A} \in \mathcal{A}rgs(\mathcal{AS}_{i+1})$ *s.t.* $Conc(\mathcal{A}) = \lceil \phi < \psi \rceil$
- $r_1 < r_2$ *in* $\mathcal{AS}_i \leftrightarrow \exists \mathcal{A} \in \mathcal{A}rgs(\mathcal{AS}_{i+1})$ *s.t.* $Conc(\mathcal{A}) = \lceil r_1 < r_2 \rceil$
- $\mathcal{A}_1 \preceq \mathcal{A}_2$ *in* $\mathcal{AS}_i \leftrightarrow \exists \mathcal{A} \in \mathcal{A}rgs(\mathcal{AS}_{i+1})$ *s.t.* $Conc(\mathcal{A}) = \lceil \mathcal{A}_1 \preceq \mathcal{A}_2 \rceil$

Definition 3.3.1 allows for either components of an object-level argumentation system to be derived from arguments in a meta-level system, or arguments in a meta-level system

to be deduced from the components of an object-level system. In broad terms, given a component Φ, of unspecified type, in an object-level argumentation system, there is an argument (that is possibly atomic) for $\lceil\Phi\rceil$ in the corresponding meta-level argumentation system; conversely, if there is an argument with conclusion $\lceil\Psi\rceil$ in a meta-level argumentation system, Ψ is a component (of whichever type) in the corresponding object-level argumentation system. Since it is not the intention of this thesis to contribute to meta-argumentation, but instead use it as a device in argument revision, the actual process of deducing meta-level arguments and object-level components will be left implicit, based on the structure of the components themselves.

Preferences between arguments can be calculated based on preferences between their non-axiom premises and/or defeasible rules. This, in principle, means that at the meta-level, arguments for the premise and rule preferences can combine in order to generate the argument preferences, either as an argument or as some form of joint attack (on an object level attack from a less-preferred to a more-preferred argument). However, characterising this would again be beyond the scope of this thesis, thus it is assumed that argument preferences are computed at the object-level, with the appropriate representation (i.e. $\mathcal{A} \prec \mathcal{B}$) being elevated as a meta-level argument, which then provides the source of the attack.

Using meta-argumentation provides several advantages for argument revision. Firstly, every component of an argumentation system becomes an argument, whether this be at the object- or a meta-level. This allows for a single method of revising arguments to be specified, which can then be applied to the meta-level arguments for ground-level components (i.e. rules, preferences and contrariness). This, in turn, allows the ground-level arguments to be revised using these components.

A further advantage of using meta-argumentation in argument revision is that it allows us to introduce a principled characterisation of the distinction between strict and defeasible rules. Recall that a rule is defeasible iff there exists some exceptional circumstance(s) in which that rule cannot be applied; otherwise, it is strict. Note that it does not need to be the

case that the source of the exception is currently true (or, in the context of an argumentation theory, there does not need to exist an acceptable argument with a conclusion that is the source of the exception) — all that needs to exist is the declaration that *if* this formula is true, the rule does not apply. Using meta-argumentation, we formally redefine the distinction between strict and defeasible rules by stating that a ground rule is strict iff its meta-level representation has no contraries at the meta-level; otherwise, it is defeasible.

Definition 3.3.2 *(Strict and defeasible rules)*
$r \in \mathcal{R}_s(\mathcal{AS}_i) \leftrightarrow \nexists \phi \in \mathcal{L}_{i+1}$ *s.t.* $\phi \in \overline{\lceil r \rceil}$
$r \in \mathcal{R}_d(\mathcal{AS}_i) \leftrightarrow \exists \phi \in \mathcal{L}_{i+1}$ *s.t.* $\phi \in \overline{\lceil r \rceil}$

In an argument revision context, this characterisation allows strict rules to be revised into defeasible rules by introducing a contrary with respect to the rule (again, without necessarily introducing the formula that is the contrary). Conversely, a defeasible rule can be revised into a strict rule by removing its contraries. The process and properties of rule revision will be described in more detail in Chapter 5.

In a broader context, representing rules as meta-arguments has a further advantage. It allows for a more precise definition of undercutting, that does not depend on Prakken's [2010] assumption that rules can be represented in the object language of an argumentation system. Instead, a defeasible rule is represented in the language of a meta-argumentation system where, per definition 3.3.2, it will have at least one contrary (in order to be defeasible). If there then exists an argument for a contrary in the ground system, the rule is undercut (possibly not successfully) by that argument. We thus formally redefine undercutting, in terms of meta-argumentation, thus:

Definition 3.3.3 *(Undercutting)*
An argument $\mathcal{A} \in \mathcal{A}rgs(\mathcal{AS}_i)$ undercuts *an argument* $\mathcal{B} \in \mathcal{A}rgs(\mathcal{AS}_i)$ *(on* $\mathcal{B}' \in Sub(\mathcal{B})$*) iff* $Conc(\mathcal{A}) \in \overline{\lceil TopRule(\mathcal{B}') \rceil}$ *in* $\mathcal{AS}_{i+1}$

In other words, an argument $\mathcal{A}$ undercuts an argument $\mathcal{B}$ (on a sub-argument $\mathcal{B}'$) if and

only if the conclusion of $\mathcal{A}$ is defined in the meta-level argumentation system as being contrary to the meta-level representation of the topmost rule in $\mathcal{B}'$.

To return to the example presented on p.43, given Definitions 3.3.1, 3.3.2 and 3.3.3 we can now encode the information as follows:

- $[r1]\ ill \Rightarrow go\_to\_doctor$ (rule at the object level)
- $have\_flu \in \overline{\lceil ill \Rightarrow go\_to\_doctor \rceil}$ (contrary at the first meta level)
- $[r2] flu\_is\_contagious \Rightarrow \lceil have\_flu \in \overline{\lceil ill \Rightarrow go\_to\_doctor \rceil} \rceil$ (rule at the second meta-level)

As can be seen, using meta-argumentation allows the justification for the exception to be represented, while also removing the need to represent rule labels in the object-level language.

To represent the set of all (meta-) argumentation systems, the notation of Weide and Dignum [2011] is employed, which describes a *tower of argumentation systems*:

Definition 3.3.4 *(Tower of Argumentation Systems)*
A tower of argumentation systems of level $1 \leq n$ *is a set* $\{\mathcal{AS}_1, \ldots, \mathcal{AS}_n\}$ *such that:*

- $\mathcal{AS}_1$ *is an argumentation system and*
- *for each* $2 \leq i \leq n$*:* $\mathcal{AS}_i$ *is a meta-argumentation system for* $\mathcal{AS}_{i-1}$

Given a tower of argumentation systems, a meta-argumentation theory is defined, again from Weide and Dignum [2011]:

Definition 3.3.5 *(Meta-argumentation theory) A Meta-Argumentation Theory (*$\mathcal{AT}$*) is a tuple* $\langle \mathcal{T}_{\mathcal{AS}}, \{\mathcal{K}_1, \ldots, \mathcal{K}_n\}, \preceq \rangle$ *such that:*

- $T_{\mathcal{AS}} = \{\mathcal{AS}_1, \ldots, \mathcal{AS}_n\}$ *is a tower of argumentation systems of level* n*, and*
- *for each* $1 \leq i \leq n$*:* $\mathcal{K}_i$ *is a knowledge base in argumentation system* $\mathcal{AS}_i$

- $\preceq$ *is an argument ordering on the set of all arguments that can be constructed from each knowledge base in each* $\mathcal{AS}$

In subsequent sections and chapters, unless made otherwise explicit, the phrase "argumentation theory" and the notation $\mathcal{AT}$ will refer to a meta-argumentation theory. The symbol Π shall be used to represent the set of all possible argumentation theories.

A core feature of ASPIC$^+$ is the ability to derive a Dung-style abstract framework from an argumentation theory. This allows structured arguments and the relations between them to be evaluated using established acceptability semantics for abstract argumentation. An abstract framework derived from an argumentation theory uses the arguments in the theory as the set of arguments, and successful attacks (defeat) in place of the attack relation.

When using meta-argumentation, an ordinary Dung-style framework cannot be derived, because of the extra levels of attack afforded — for instance, attacks on attacks (through an attack on a preference of contrariness relation). Modgil and Prakken [2010] introduced the idea of attacks on attacks in abstract argumentation, through the use of Extended Argumentation Frameworks (EAFs). Weide and Dignum [2011] provide a definition of how to obtain an EAF from an ASPIC$^+$ meta-argumentation theory, which is employable in the present work:

Definition 3.3.6 *(Structured EAF)*
$EAF = \langle \mathcal{A}, \mathcal{C}, \mathcal{D} \rangle$ *is a bounded hierarchical EAF, where* $\mathcal{A} = Args(\mathcal{AT})$ *and:*

- $(\mathcal{A}_1, \mathcal{A}_2) \in \mathcal{C}$ *if* $\mathcal{A}_1$ *attacks* $\mathcal{A}_2$.
- *if* $\exists \mathcal{A}_1, \mathcal{A}_2 \in \mathcal{A}$ *s.t* $Conc(\mathcal{A}_2) = \lceil r : \varphi_1, \ldots, \varphi_n \Rightarrow \varphi \rceil$ *and* $Conc(\mathcal{A}_1) \in \overline{Conc(\mathcal{A}_2)}$, *then* $(\mathcal{A}_1, \mathcal{A}_2) \in \mathcal{C}$ *and* $\forall \mathcal{A}_3 \in \mathcal{A}$ *s.t.* $r \in DefRules(\mathcal{A}_3)$, $(\mathcal{A}_1, \mathcal{A}_3) \in \mathcal{C}$
- $\forall (\mathcal{A}_1, \mathcal{A}_2) \in \mathcal{C}$, *if* $\exists \mathcal{A}_3 \in \mathcal{A}$ *s.t.* $Conc(\mathcal{A}_3) \in \overline{\lceil Conc(\mathcal{A}_1) \in \overline{Conc(\mathcal{A}_2)} \rceil}$, *then* $(\mathcal{A}_3, (\mathcal{A}_1, \mathcal{A}_2)) \in \mathcal{D}$ *and* $\forall \mathcal{A}_4 \in \mathcal{A}$ *s.t.* $\mathcal{A}_4 \in Sub(\mathcal{A}_2)$, $(\mathcal{A}_3, (\mathcal{A}_1, \mathcal{A}_4)) \in \mathcal{D}$
- $\forall (\mathcal{A}_1, \mathcal{A}_2) \in \mathcal{C}$, *if* $\exists \mathcal{A}_3 \in \mathcal{A}$ *s.t.* $Conc(\mathcal{A}_3) = \lceil \mathcal{A}_1 \prec \mathcal{A}_2 \rceil$, *then* $(\mathcal{A}_3, (\mathcal{A}_1, \mathcal{A}_2)) \in \mathcal{D}$ *and* $\forall \mathcal{A}_4 \in \mathcal{A}$ *s.t.* $\mathcal{A}_4 \in Sub(\mathcal{A}_2)$, $(\mathcal{A}_3, (\mathcal{A}_1, \mathcal{A}_4)) \in \mathcal{D}$

The second and third parts of definition 3.3.6 introduce not only an attack on the attack between $\mathcal{A}_1$ and $\mathcal{A}_2$, but also the attack by $\mathcal{A}_1$ on any arguments in which $\mathcal{A}_2$ is a sub-argument.

In the ASPIC$^+$ framework, when an abstract framework is derived, it uses the notion of *defeat* in place of attack (with a defeat being an attack which is successful, w.r.t. preferences). However, when using meta-argumentation, preferences and attacks become arguments which can themselves be attacked. An EAF, therefore, reverts to the use of attack, with defeat being parametrised with respect to preference, specified by a set S of arguments [Modgil and Prakken, 2010]: $Y \rightarrow^S X$ iff $(Y, X) \in \mathcal{C}$ and $\nexists Z \in S$ s.t. $(Z, (Y, X)) \in \mathcal{D}$. In other words, an argument Y defeats an argument X iff there exist an attack on X by Y and there is no argument Z in S that attacks the attack.

Acceptability in EAFs differs from that found in a standard Dung AF. This definition is also taken from [Modgil and Prakken, 2010]:

Definition 3.3.7 *(EAF acceptability)*
Given an Extended Argumentation Framework $EAF = \langle \mathcal{A}, \mathcal{C}, \mathcal{D} \rangle$*, let* $S \subseteq \mathcal{A}$*. Let* $\mathcal{N}_S = \{X_1 \rightarrow^S Y_1, \ldots, X_n \rightarrow^S Y_n\}$ *where for* $i = 1 \ldots n$*,* $X_i \in S$*. Then* $\mathcal{N}_S$ *is a reinstatement set for* $A \rightarrow^S B$ *iff* $A \rightarrow^S B \in \mathcal{N}_S$*, and* $\forall X \rightarrow^S Y \in \mathcal{N}_S, \forall Y'$ *s.t.* $(Y', (X, Y)) \in \mathcal{D}$*,* $\exists X' \rightarrow^S Y' \in \mathcal{N}_S$

X is acceptable w.r.t. $S \subseteq \mathcal{A}$ iff $\forall Y$ s.t. $Y \rightarrow^S X$, there is a reinstatement set for some $Z \rightarrow^S Y$.

Given definition 3.3.7, extensions are defined in the same way as for a standard Dung AF [Dung, 1995].

3.4 Summary

This chapter has introduced two important foundations upon which the rest of the thesis will build. The ASPIC$^+$ framework of [Prakken, 2010] was first introduced, before going

on to define meta-level extensions to it. The meta-level extensions elevate object-level components (conflict, rules and preferences) to meta-level arguments, and differs from previous work on meta-argumentation in dialogue [Weide and Dignum, 2011] by mandating that every object-level component be represented by a corresponding meta-level argument. Definitions were then provided from [Modgil and Prakken, 2010] for the derivation, and evaluation of an extended argumentation framework from a meta-argumentation theory.

Chapter 4

Dialogue framework: SPD

4.1 Introduction

In this chapter, a dialogue framework is presented in which the participants use a system of structured argumentation to formulate locutions. The framework, named SPD (Simple Persuasion Dialogue), describes a communication language and protocol for a persuasion dialogue, with the former being specified on the basis of the ASPIC$^+$ framework.

Weide and Dignum [2011] previously specified a dialogue protocol based on the ASPIC$^+$ framework, defined in terms of meta-argumentation systems and designed with the intention of allowing participants to reason and argue about preferences between arguments. The framework was specified in terms of towers of meta-argumentation systems; a meta-argumentation system being an argumentation system that describes another argumentation system, and a tower being a set of these, where the system $\mathcal{AS}_n$ describes the system $\mathcal{AS}_{n-1}$.

Weide and Dignum's protocol has two significant drawbacks; the first is that its $claim$ locution is defined over ASPIC$^+$ *arguments* and not formulae of the logical language. Thus if an agent possess an argument for some conclusion ϕ, it must, from the definition of an argument (see [Prakken, 2010], as outlined in Chapter 2), claim its entire argument for it. This presents two disadvantages: firstly, it can result in relatively trivial dialogues, where

one participant claims an argument and his opponent claims another argument that defeats it — there is no scope to claim only the conclusion and initiate a claim-challenge-defend process as found in other protocols, such as those specified by Walton and Krabbe [1995].

The second drawback is the inability of the protocol to generate a new argumentation theory that is not simply the combination of parts of those from which the participants' formulated their locutions. By allowing arguments to be broken up and reconstructed as part of a dialogical process, we can end up with what are essentially new arguments, with new interactions (i.e. attacks and defeat).

This chapter proceeds as follows: first, the concepts of shared and personal argumentation theories, and their role in a dialogue, are introduced; a communication language, inspired by that of Weide and Dignum [2011], is then presented, with key differences highlighted. Finally, commitment and structural rules are specified, with the former dictating how the commitments of dialogue participants are updated, and the latter governing how a dialogue begins, progresses and terminates.

4.2 Personal and shared argumentation theories

Before specifying the dialogue framework itself, the concept of personal and shared argumentation theories is first introduced.

4.2.1 Personal argumentation theory

Each participant engaging in an SPD dialogue will possess a *personal argumentation theory*, from which they will determine belief for the purposes of advancing locutions in the dialogue. For a participant α, their personal argumentation theory is represented as $\mathcal{PAT}_\alpha$. Each participant's personal argumentation theory will contain at least one argumentation system. Furthermore, participants will share a common language at each meta-level, i.e. for two participants α and β, the language of $\mathcal{AS}_i$ in $\mathcal{PAT}_\alpha$ is the same as the language of $\mathcal{AS}_i$ in $\mathcal{PAT}_\beta$. However, it is not required that each personal argumentation theory extend

to the same meta-level — so for instance, $\mathcal{PAT}_\alpha$ may contain n argumentation systems, while $\mathcal{PAT}_\beta$ may contain m, with $n \neq m$. However, for the purposes of specifying SPD, we assume that the topmost language in all personal argumentation theories is $\mathcal{L}_n$, which subsumes all lower-level languages.

From its personal argumentation theory, a participant derives its beliefs. For the purposes of this work, we shall assume that all participants are sceptical and thus will use grounded semantics as the method of evaluating the abstract framework. Grounded semantics are also chosen because they yield a unique extension and determining the nature of belief based on multi-extensional semantics is beyond the scope of this thesis.

Definition 4.2.1 *(Participant beliefs)*
Given a personal argumentation theory $\mathcal{PAT}_\alpha$ for a participant α, α's beliefs are $B_\alpha = Cl_{R_s}(\{Conc(A) \mid \mathcal{A} \in E_G(\mathcal{PAT}_\alpha)\})$, where Cl_{R_s} is the strict rule closure operator (Chapter 3, p.42, Definition 3.2.8) and $E_G(\mathcal{PAT}_\alpha)$ is the grounded extension of the abstract framework derived from $\mathcal{PAT}_\alpha$.

That is, a participant believes the conclusion of all arguments in the grounded extension of the framework derived from its personal argumentation theory, as well as anything that can be inferred from it under strict rule application.

Private arguments and beliefs

A participant in a dialogue may wish to avoid uttering certain information they believe. For instance, in a real domain, a politician may wish to keep private an opinion that does not garner much public support, or a researcher might not wish to reveal unpublished results from a study. Alternatively, at a more abstract level, if we were using a value-based argumentation framework [Bench-Capon and Dunne, 2002], if an argument $\mathcal{A}$ promotes some negative value v, then a participant may wish to keep it private.

While it is beyond the scope of this thesis to fully explore the nature of what constitutes private arguments and, by extension, beliefs, and the criteria used in determining them,

we still capture the concept for use in SPD, and to illustrate an application of argument revision in Chapter 6. For now, we assume that a participant in a dialogue has a set of acceptable arguments $P \subseteq E_G(\mathcal{PAT})$ which, using some criteria, they have determined to be "private".

If an argument is private, then any arguments in which it is a sub-argument should also be marked as private. We capture this through the definition of a closure operator, Cl_p.

Definition 4.2.2 *(Private argument closure)*
Let P_α be a set of private arguments. $Cl_p(P_\alpha)$ is the smallest set such that:

- $P_\alpha \subseteq Cl_p(P_\alpha)$
- $\forall \mathcal{A} \in Cl_p(P_\alpha), \mathcal{A}' \in E_G(\mathcal{PAT}_\alpha)$, *if* $\mathcal{A} \in Sub(\mathcal{A}')$, $\mathcal{A}' \in Cl_p(P_\alpha)$

Given the closure of a participant's private arguments, we now define their *communicable beliefs*, which is a set of all conclusions of non-private arguments.

Definition 4.2.3 *(Communicable beliefs)*
Let P_α be a set of private arguments. The ***communicable beliefs*** *of α are:*
$B^c_\alpha = Cl_{R_s}(\{Conc(A) \mid \mathcal{A} \in E_G(\mathcal{PAT}_\alpha)\} \backslash Cl_p(P_\alpha))$

Note that it is not necessarily the case that $B^c_\alpha = B_\alpha \backslash \{Conc(\mathcal{A}) \mid \mathcal{A} \in Cl_p(P_\alpha)\}$. This is because there may exist more than one argument with the same conclusion in $E_G(\mathcal{PAT}_\alpha)$, but when deriving belief, the distinction between arguments is lost and only one instance of the conclusion is passed into B_α. Thus in order to maintain all non-private conclusions, we need to return to the arguments themselves to derive the communicable beliefs.

4.2.2 Shared argumentation theory

During a dialogue, the participants will construct a *shared argumentation theory*, denoted as $\mathcal{AT}_D$. The shared theory will contain as many argumentation systems and knowledge

bases as are required to reflect the current meta-level of the dialogue; for instance, if an entire dialogue is conducted only at the object-level, the shared theory will contain only one argumentation system and associated knowledge base. However, if the dialogue contains elements at the meta-level, it would contain two argumentation systems and knowledge bases (at the meta-meta-level it would contain three and so on).

The shared theory has two purposes:

- To provide a monological account of the dialogue
- To provide an assessment of the acceptability of the arguments advanced in the dialogue

Providing a monological account of the dialogue allows the arguments expressed to be analysed and evaluated post-dialogue. This provides an overall view of everything that was advanced and, ultimately, the outcome.

Assessing the acceptability of arguments during the dialogue allows the participants to assess what their next locution should be. For instance, if a participant finds that an argument they previously advanced is defeated in the shared system, they can either offer a defence for it, or be forced into retracting it by their opponent. This process will be explained further by the structural rules of the dialogue (section 4.5).

The construction of the knowledge bases of the argumentation systems in $\mathcal{AT}_D$ is closely connected to commitment, and the structure of the dialogue. This will be elaborated in the commitment and structural rules (sections 4.4 and 4.5 respectively).

4.3 Communication language

In this section, the communication language for SPD is specified. The communication language defines the valid locutions, and the content thereof, that a participant can make.

Definition 4.3.1 *(Communication language)*

A communication language $\mathcal{L}_C$ for a language $\mathcal{L}_n$ is:

- $\forall \Phi \in 2^{\mathcal{L}_n}$: $claim(\Phi)$, $why(\Phi)$, $concede(\Phi)$, $resolve(\Phi) \in \mathcal{L}_C$
- $\forall \Phi \in 2^{\mathcal{L}_n}, \Psi \in 2^{\mathcal{L}_n}$ *s.t.* $\Phi \cap \Psi = \emptyset$, $retract(\Phi, \Psi) \in \mathcal{L}_C$

The content of each locution is a set of formulae of the ASPIC$^+$ language $\mathcal{L}_n$, with the exception of $retract$ which has two sets as its content — the first is the formula to be retracted, while the latter is a (possibly empty) set of justifications for the retraction. The concept of justified retractions will be explained further in section 4.4.2.

Locutions are advanced as part of a move. Moves contain an identifier, the participant, the locution and a timestamp.

Definition 4.3.2 *(Dialogue move)*
A dialogue move is a tuple, $m = \langle id, pl, loc, t \rangle$ *where:*

- $id \in \mathbb{N}$ *the ID of the move*
- $pl \in \mathcal{P}$*, the participant*
- $loc \in \mathcal{L}_C$*, the locution, including its content*
- $t \in \mathbb{N}$*, the target of the move*

For notational convenience, each element of a move is provided with a corresponding function which returns the value of that element in the move. For instance, for a move $m_1 = \langle 1, \alpha, claim(\phi), 0 \rangle$:

- $id(m_1) = 1$
- $pl(m_1) = \alpha$
- $loc(m_1) = claim(\phi)$
- $t(m_1) = 0$

A dialogue consists of a topic, a communication language, a set of participants and a set of moves. To represent a dialogue, we use notation similar to that of Reed [1998] and Weide and Dignum [2011]:

Definition 4.3.3 *(Dialogue)*
A dialogue is a tuple $\mathcal{D} = \langle \tau, \mathcal{L}_C, \mathcal{P}, \mathcal{M} \rangle$, where:

- *$\tau \in \mathcal{L}_n$ is the topic of the dialogue*
- *$\mathcal{L}_C$ is a communication language on the basis of definition 4.3.1*
- *$\mathcal{P}$ is a set of participants*
- *$\mathcal{M}$ is a set of moves on the basis of definition 4.3.2*

4.4 Commitment and commitment rules

As a dialogue progresses, participants incur commitment to formulae they claim and concede. In this section, the structure of a participant's commitment store and the rules governing how commitment is incurred are specified.

4.4.1 Commitment store

A commitment store is a representation of everything to which a participant has become committed during a dialogue. Commitment is incurred when a participant makes a claim and when they concede a claim made by an opponent.

Definition 4.4.1 *(Commitment store)*
The commitment store of a participant $\alpha \in \mathcal{P}$ at time t in a dialogue is C_α^t, such that $C_\alpha^t \subseteq \mathcal{L}_n$

A commitment store is non-monotonic, in that a participant can retract commitment to a formula; however, it is possible that their remaining commitments still allow a retracted

formula to be inferred. To determine whether or not this is the case, we define a closure operator over commitment stores.

Definition 4.4.2 *(Commitment closure)*

Let C^t_α be a commitment store. $Cl_C(C^t_\alpha)$ is the smallest set such that:

- $C^t_\alpha \subseteq Cl_C(C^t_\alpha)$
- *if* $\lceil \varphi_1, \ldots, \varphi_n \rightsquigarrow \varphi \rceil \in Cl_C(C^t_\alpha)$ *then if* $\{\varphi_1, \ldots, \varphi_n\} \subseteq Cl_C(C^t_\alpha)$, $\varphi \in Cl_C(C^t_\alpha)$

A commitment store is closed iff $C^t_\alpha = Cl_C(C^t_\alpha)$.

From the definition of commitment store closure, notions of direct and indirect consistency are now defined. A commitment store is consistent iff it does not contain two formulae where one is a contrary of the other (with the contrary relation itself represented as a formula in the commitment store).

Definition 4.4.3 *(Direct consistency)*

A commitment store C_α is directly consistent iff for any $\phi, \phi' \in C_\alpha$, $\{\phi, \phi', \lceil \phi' \in \overline{\phi} \rceil\} \not\subseteq C_\alpha$.

Indirect inconsistency requires that the conditions for direct consistency be fulfilled in the closure of the commitment store.

Definition 4.4.4 *(Indirect consistency)*

A commitment store C_α is indirectly consistent iff for any $\phi, \phi' \in Cl_C(C_\alpha)$, $\{\phi, \phi', \lceil \phi' \in \overline{\phi} \rceil\} \not\subseteq Cl_C(C_\alpha)$.

4.4.2 Commitment rules

Having now defined what constitutes a commitment store, how it is closed and the conditions under which it becomes inconsistent, the commitment rules of SPD can be specified. The commitment rules govern how a participant incurs commitment as a dialogue progresses, based on their locutions. The commitment rules also direct the construction of the shared argumentation theory, $\mathcal{AT}_D$. For notational convenience, $\mathcal{AS}^i_D$ will be used as

shorthand for "the argumentation system containing the meta-level language to which this formula belongs".

The first commitment rule states that when a participant claims a formula, they become committed to that formula. Additionally, the formula is added to the knowledge base of the corresponding meta-level argumentation system in $\mathcal{AS}_D$.

C1 if $loc(m_i) = claim_\alpha(\Phi)$ then:

- $\mathcal{C}_\alpha^i = \mathcal{C}_\alpha^{i-1} \cup \Phi$
- Given $\Phi = \Phi_1 \cup \Phi_2$ such that $\Phi_1 \cap \Phi_2 = \emptyset$, $\forall \phi \in \Phi$, if $\exists \psi, \lceil \phi \in \overline{\psi} \rceil \in \mathcal{C}l_c(C_\alpha^{i-1})$ $\phi \in \Phi_1$, else $\phi \in \Phi_2$. Then $\mathcal{K}_a(\mathcal{AS}_D^i) = \mathcal{K}_a(\mathcal{AS}_D^{i-1}) \cup \Phi_1$ and $\mathcal{K}_p(\mathcal{AS}_D^i) = \mathcal{K}_p(\mathcal{AS}_D^{i-1}) \cup \Phi_2$.

The second effect is stating that if this participant has previously claimed a formula of which the newly-claimed formula is a contrary (and the contrary relation), the new formula is added to $\mathcal{AT}_D$ as an assumption. This will be explained further in the specification of the structural rules, section 4.5.

As well as updating the knowledge base of an argumentation system in $\mathcal{AT}_D$, if a claimed formula is at the meta-level, the object-level component (rule, preference, contrariness) that it represents will be added to the corresponding object-level argumentation system.

The second commitment rule relates to conceding a set of formulae to an opponent.

C2 if $loc(m_i) = concede_\alpha(\Phi)$, then $\mathcal{C}_\alpha^i = \mathcal{C}_\alpha^{i-1} \cup \Phi$

There is no requirement for $concede$ to update $\mathcal{AT}_D$, because the formulae will have already been added as a result of a $claim$ from the opponent.

The final commitment rule describes the process of retraction. A retraction is, in its simplest form, the opposite of a claim, in that it takes a set of formulae out of the commitment store. However, it can also add new formulae, in order to justify the retraction.

C3 if $loc(m_i) = retract_\alpha(\Phi, \Psi)$ then:

- $C^i_\alpha = (C^{i-1}_\alpha \backslash \Phi) \cup \Psi$
- $\forall \psi \in \Psi$, if $\psi \notin \mathcal{K}(\mathcal{AS}^i_D)$, $\mathcal{K}_a(\mathcal{AS}^i_D) = \mathcal{K}_a(\mathcal{AS}^{i+1}_D) \cup \{\psi\}$

A retraction does not remove the formulae from the knowledge base of an argumentation system in $\mathcal{AT}_D$, because of the deductive monotonicity of $\mathcal{AT}_D$. Furthermore, a retract locution does not remove formulae from the *closure* of the commitment store — it is entirely possible that a retracted formula can still be inferred from what remains, which means the participant should either retract additional formulae, or face being forced into a further retraction by their opponent.

The second parameter of a $retract$ is a set of formulae that are used to justify the retraction, if no justification already exists in the participant's commitment store (as a result of a concession). A justification involves incurring commitment to a formula that is a contrary of the retracted formula, along with the associated contrariness relation (e.g. $\{\phi, \ulcorner \phi \in \overline{\psi} \urcorner\}$ would justify a retraction of ψ). Note that there is no requirement for a retraction to be justified, so even if one does not already exist in the participant's commitment store, they do not have to provide one in the $retract$ locution.

To distinguish between justified and unjustified retractions, we formally define what constitutes justification.

Definition 4.4.5 *(Justified retraction)*
A retraction $retract(\Phi, \Psi)$ *at step* t *is* ***justified*** *iff:* $\forall \phi \in \Phi$*, then for some* ψ*:* $\{\psi, \ulcorner \psi \in \overline{\phi} \urcorner\} \subseteq Cl_C(C^{t-1} \cup \Psi)$.

Remark 7 *We consider the closure of* $C^{t-1} \cup \Psi$ *and not* C^t *to ensure that if a justification is provided in the* $retract$ *locution, it is considered in the evaluation of justification.*

Remark 8 *Where a retraction is unjustified, we use a superscript* u *on the locution:* $retract^u(\Phi, \Psi)$.

In other words, a retraction is justified if a justification is present in the closure of the commitment store following the $retract$. Formulae expressed in a justification as part of a retraction are added to the assumptions set of the knowledge base in the relevant argumentation systems in $\mathcal{AT}_D$ because there is a (possibly erroneous; see Chapter 6) assumption of honesty between participants. Thus, if a participant claims a formula, they are seen as believing it; if they then claim a contrary formula to justify retracting the original claim, they are seen as having updated their beliefs to *assume* the contrary is true (and hence why it is the case their belief no longer holds).

A present constraint in the dialogue framework is that for a retraction to be justified, all formulae in the retraction must be justified (i.e. have appropriate contraries in the closure of the commitment store or justification parameter in the locution). It will be left to future work to investigate the concept of "partial justification", where justification is provided for only a proper subset of Φ.

The use of justified retraction will be explained more fully in section 4.5, but generally speaking if a participant does not justify a retraction, they cannot reclaim the retracted formula later in the dialogue, even if all other conditions allow for it.

4.5 Structural rules

The structural rules in a dialogue specify the protocol, dictating whose turn it is and what locutions are valid. The structural rules of SPD are inspired by those of RPD_0 [Walton and Krabbe, 1995]. For simplicity, it will be assumed the dialogue is taking place with only two participants, $\mathcal{P} = \{\alpha, \beta\}$.

The first structural rule describes valid moves following a move with a $claim$ locution. Following a claim by one participant, the second participant must either question it, (counter-)claim a contrary or contradictory formula, concede the claimed formula or, if the claim has caused or failed to resolve an inconsistency, force a resolution.

R1 if $loc(m_i) = claim_\alpha(\Phi)$, there must $\nexists m' \in \mathcal{M}$ s.t.$pl(m') = \alpha$ and $loc(m') =$

$retract^u(\Phi, \emptyset)$ and $loc(m_{i+1})$ must be either:

- $why_\beta(\Phi)$,
- $claim_\beta(\{\psi, \lceil \psi \in \overline{\phi} \rceil\})$, for some $\phi \in \Phi$, with $\mathcal{K}_p(\mathcal{AS}_D^{i+1}) = \mathcal{K}_p(\mathcal{AS}_D^i) \cup \{\psi, \lceil \psi \in \overline{\phi} \rceil\}$
- $concede_\beta(\Phi)$
- $resolve_\beta(\Phi)$

The initial constraint placed on the content of a $claim$ means that a set of formulae can be claimed iff it has not been unjustifiably retracted at an earlier point in the dialogue. We again place a constraint that only a full claim (i.e. all the formulae within it) can be conceded; "partial concession", where only a proper subset of the claimed formulae are conceded, will be left for future investigation.

The second structural rule describes valid moves following a move containing a why locution. The participant whose claim is questioned can either re-assert the formulae, justify them with rules, or retract them.

R2 if $loc(m_i) = why_\beta(\Phi)$, then $loc(m_{i+1})$ must be either:

- $claim_\alpha(\Phi)$,
- $claim_\alpha(\Phi')$, where $\Phi' = \bigcup_{\phi \in \Phi} \{\phi_1, \ldots, \phi_n, \lceil \phi_1, \ldots, \phi_n \Rightarrow \phi \rceil\}$
- $retract_\alpha(\Phi, \Psi)$

When a participant concedes, its opponent can either force it to resolve either an inconsistency in its commitment store, or a lack of defence of a previous claim. Alternatively, any other move can be made provided the dialogue has not terminated.

R3 if $loc(m_i) = concede_\alpha(\Phi)$ then $loc(m_{i+1})$ must be either:

- $resolve_\beta(\Phi')$ if for $\Phi' \subseteq \Phi$, $Cl_C(C_\alpha^i)$ is inconsistent w.r.t. Φ', or $\forall \phi \in \Phi'$, all arguments for ϕ are defeated in $\mathcal{AT}_D$

- any other non-$resolve$ move by β, except in the conditions of **R5**

The second condition of **R3** allows for an inconsistency or lack of defence to go unchallenged. While all examples provided in this thesis will demand such resolutions, there is no technical limitation to allowing them to pass and hence the ability is provided in the protocol.

The fourth rule describes the response to a participant being forced to resolve its commitment store. This takes the form of a retraction of the formulae at the source of the inconsistency, while also providing a justification for the retraction, unless the participant chooses not to provide one, or a justification has already been given earlier in the dialogue.

R4 if $loc(m_i) = resolve_\alpha(\Phi)$ then $loc(m_{i+1})$ must be: $retract_\beta(\Phi, \Psi)$.

The final rule describes the termination conditions. A dialogue terminates when either the proponent (α) retracts the topic of the dialogue, or the opponent (β) concedes it. Note that since concession of the topic by the proponent terminates the dialogue, there is no requirement for his commitment store to be resolved (i.e. he does not need to perform a stability adjustment).

R5 a dialogue terminates at move m_i if either:

- $loc(m_i) = concede_\beta(\tau)$
- $loc(m_i) = retract_\alpha(\tau)$.

4.6 Running example

In this section, an example will be presented that will be used not only to illustrate SPD, but also techniques presented in later chapters.

Consider two participants, $\mathcal{P} = \{\alpha, \beta\}$. Each possesses a personal argumentation theory. The construction of $\mathcal{PAT}_\alpha = \langle\{\mathcal{AS}_1, \mathcal{AS}_2, \mathcal{AS}_3\}, \{K_1, \mathcal{K}_2, \mathcal{K}_3\}\rangle$ is as follows

(where the letters represent formulae of the language $\mathcal{L}_n$):

$\mathcal{K}_n(\mathcal{AS}_1) = \{w_1, z\}$

$$\mathcal{K}_p(\mathcal{AS}_1) = \left\{ \begin{array}{lll} a_1, & a_2, & b_1, \\ c_1, & c_2, & e_1, \\ f_1, & g_1, & h_1, \\ i & & \end{array} \right\}$$

$$\mathcal{R}_d(\mathcal{AS}_1) = \left\{ \begin{array}{ll} a_1, a_2 \Rightarrow_{r1} a, & b_1 \Rightarrow_{r2} b, \\ c_1, c_2 \Rightarrow_{r3} c, & e_1 \Rightarrow_{r4} e, \\ f_1 \Rightarrow_{r5} f, & g_1 \Rightarrow_{r6} g, \\ h_1 \Rightarrow_{r7} h, & w_1 \Rightarrow_{r8} w, \\ a_2 \Rightarrow_{r9} u & \end{array} \right\}$$

$$\mathcal{K}_p(\mathcal{AS}_2) = \left\{ \begin{array}{lll} \lceil a_1, a_2 \Rightarrow_{r1} a \rceil, & \lceil b_1 \Rightarrow_{r2} b \rceil, & \lceil c_1, c_2 \Rightarrow_{r3} c \rceil, \\ \lceil e_1 \Rightarrow_{r4} e \rceil, & \lceil f_1 \Rightarrow_{r5} f \rceil, & \lceil g_1 \Rightarrow_{r6} g \rceil, \\ \lceil h_1 \Rightarrow_{r7} h \rceil, & \lceil w_1 \Rightarrow_{r8} w \rceil, & \lceil a_2 \Rightarrow_{r9} u \rceil, \\ \lceil b \in \overline{a} \rceil, & \lceil f \in \overline{d} \rceil, & \lceil g \in \overline{d} \rceil, \\ \lceil h \in \overline{g} \rceil, & \lceil h \in \overline{i} \rceil, & \lceil u \in \overline{e_1} \rceil, \\ \lceil y \in \overline{z} \rceil, & \lceil x \in \overline{w_1} \rceil, & \lceil a_2 \in \overline{e_1} \rceil, \\ \lceil x \in \overline{a_1} \rceil & & \end{array} \right\}$$

$\mathcal{K}_p(\mathcal{AS}_3) = \{\lceil c \in \overline{\lceil b_1 \Rightarrow_{r2} b \rceil} \rceil, \lceil y \in \overline{\lceil a_1, a_2 \Rightarrow_{r1} a \rceil} \rceil\}$

From these argumentation systems, α can construct the following arguments:

$\mathcal{A}_1 : w_1$	$\mathcal{A}_2 : z$	$\mathcal{A}_3 : a_1$
$\mathcal{A}_4 : a_2$	$\mathcal{A}_5 : b_1$	$\mathcal{A}_6 : c_1$
$\mathcal{A}_7 : c_2$	$\mathcal{A}_8 : e_1$	$\mathcal{A}_9 : f_1$
$\mathcal{A}_{10} : g_1$	$\mathcal{A}_{11} : h_1$	$\mathcal{A}_{12} : i$

$\mathcal{A}_{13} : \mathcal{A}_1 \Rightarrow_{r8} w$ $\mathcal{A}_{14} : \mathcal{A}_3, \mathcal{A}_4 \Rightarrow_{r1} a$ $\mathcal{A}_{15} : \mathcal{A}_5 \Rightarrow_{r2} b$

$\mathcal{A}_{16} : \mathcal{A}_6, \mathcal{A}_7 \Rightarrow_{r3} c$ $\mathcal{A}_{17} : \mathcal{A}_8 \Rightarrow_{r4} e$ $\mathcal{A}_{18} : \mathcal{A}_9 \Rightarrow_{r5} f$

$\mathcal{A}_{19} : \mathcal{A}_{10} \Rightarrow_{r6} g$ $\mathcal{A}_{20} : \mathcal{A}_{11} \Rightarrow_{r7} h$ $\mathcal{A}_{21} : \mathcal{A}_4 \Rightarrow_{r9} u$

$\mathcal{A}'_1 : \lceil a_1, a_2 \Rightarrow_{r1} a \rceil$ $\mathcal{A}'_2 : \lceil b_1 \Rightarrow_{r2} b \rceil$ $\mathcal{A}'_3 : \lceil c_1, c_2 \Rightarrow_{r3} c \rceil$

$\mathcal{A}'_4 : \lceil e_1 \Rightarrow_{r4} e \rceil$ $\mathcal{A}'_5 : \lceil f_1 \Rightarrow_{r5} f \rceil$ $\mathcal{A}'_6 : \lceil g_1 \Rightarrow_{r6} g \rceil$

$\mathcal{A}'_7 : \lceil h_1 \Rightarrow_{r7} h \rceil$ $\mathcal{A}'_8 : \lceil w_1 \Rightarrow_{r8} w \rceil$ $\mathcal{A}'_9 : \lceil a_2 \Rightarrow_{r9} u \rceil$

$\mathcal{A}'_{10} : \lceil b \in \overline{a} \rceil$ $\mathcal{A}'_{11} : \lceil f \in \overline{d} \rceil$ $\mathcal{A}'_{12} : \lceil g \in \overline{d} \rceil$

$\mathcal{A}'_{13} : \lceil h \in \overline{g} \rceil$ $\mathcal{A}'_{14} : \lceil h \in \overline{i} \rceil$ $\mathcal{A}'_{15} : \lceil y \in \overline{z} \rceil$

$\mathcal{A}'_{16} : \lceil x \in \overline{w_1} \rceil$ $\mathcal{A}'_{17} : \lceil a_2 \in \overline{e_1} \rceil$ $\mathcal{A}'_{18} : \lceil x \in \overline{a_1} \rceil$

$\mathcal{A}''_1 : \lceil c \in \overline{\lceil b_1 \Rightarrow_{r2} b \rceil} \rceil$ $\mathcal{A}''_2 : \lceil y \in \overline{\lceil a_1, a_2 \Rightarrow_{r1} a \rceil} \rceil$

For clarity, we will consider only object-level arguments when examining argument acceptability and belief. Only one meta-argument, $\mathcal{A}'_2$ is not acceptable, due to a successful attack by $\mathcal{A}_{16}$.

The grounded extension from the abstract framework derivable from $\mathcal{PAT}_\alpha$ is:

$$E_G(\mathcal{PAT}_\alpha) = \left\{ \begin{array}{lll} \mathcal{A}_1 : w_1, & \mathcal{A}_2 : z, & \mathcal{A}_3 : a_1, \\ \mathcal{A}_4 : a_2, & \mathcal{A}_5 : b_1, & \mathcal{A}_6 : c_1, \\ \mathcal{A}_7 : c_2, & \mathcal{A}_9 : f_1, & \mathcal{A}_{10} : g_1, \\ \mathcal{A}_{11} : h_1, & \mathcal{A}_{13} : \mathcal{A}_1 \Rightarrow_{r7} w, & \mathcal{A}_{14} : \mathcal{A}_3, \mathcal{A}_4 \Rightarrow_{r1} a \\ \mathcal{A}_{16} : \mathcal{A}_6, \mathcal{A}_7 \Rightarrow_{r3} c, & \mathcal{A}_{18} : \mathcal{A}_9 \Rightarrow_{r5} f, & \mathcal{A}_{20} : \mathcal{A}_{11} \Rightarrow_{r7} h, \\ \mathcal{A}_{21} : \mathcal{A}_4 \Rightarrow_{r9} u & & \end{array} \right\}$$

Assume that $P_\alpha = \{\mathcal{A}_9\}$; $Cl_P(P_\alpha) = \{\mathcal{A}_9, \mathcal{A}_{18}\}$. Thus α's communicable beliefs are,

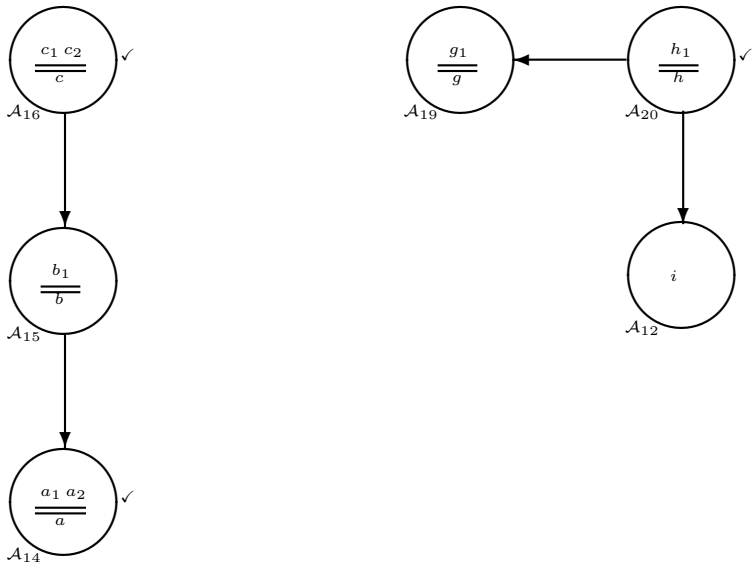

Figure 4.1: Framework from $\mathcal{PAT}_\alpha$

$$B^c_\alpha = \left\{ \begin{array}{lll} w_1, & w, & z, \\ a_1, & a_2, & a, \\ b_1, & c_1, & c_2, \\ c, & g_1, & h_1, \\ h, & u & \end{array} \right\}$$

Despite having acceptable arguments for both f_1 and f, the argument for f_1 is private and thus renders both it, and the argument for f incommunicable.

The abstract framework derivable from $\mathcal{PAT}_\alpha$ is shown in Figure 4.1, with islands (arguments that neither attack nor are attacked) removed for clarity. Arguments marked with a tick (✓) are acceptable.

The purpose of this example will be to demonstrate an argument revision process being performed by α. Therefore only minimal knowledge bases and sets of rules are provided for β, sufficient to allow the dialogue to proceed non-trivially, but without obscuring the illustration of the concepts in subsequent chapters.

The construction of $\mathcal{PAT}_\beta = \langle \{\mathcal{AS}_1, \mathcal{AS}_2, \mathcal{AS}_3\}, \{K_1, \mathcal{K}_2, \mathcal{K}_3\} \rangle$ is as follows:

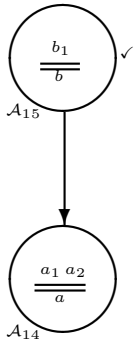

Figure 4.2: Framework from $\mathcal{PAT}_\beta$

$\mathcal{K}_p(\mathcal{AS}_1) = \{a_1, a_2, b_1, d_1, d_2\}$
$\mathcal{R}_d(\mathcal{AS}_1) = \{(a_1, a_2 \Rightarrow_{r1} a), (b_1 \Rightarrow_{r2} b), (d_1, d_2 \Rightarrow_{r10} d)\}$

$\mathcal{K}_p(\mathcal{AS}_2) = \{\lceil b \in \overline{a} \rceil\}$

$\mathcal{K}_p(\mathcal{AS}_3) = \emptyset$

$\mathcal{K}_p(\mathcal{AS}_4) = \{\lceil d \in \overline{\lceil c \in \overline{\lceil b_1 \Rightarrow_{r2} b \rceil} \rceil} \rceil\}$

For clarity in notation, arguments that can be constructed from $\mathcal{PAT}_\beta$ that are also in present in $\mathcal{PAT}_\alpha$ are omitted below, and arguments that are not present in $\mathcal{PAT}_\alpha$ (acceptable or otherwise) will be numbered starting from where α left off, at each level.

$\mathcal{A}_{22} : d_1$ $\quad$ $\mathcal{A}_{23} : d_2$
$\mathcal{A}_{24} : \mathcal{A}_{14}, \mathcal{A}_{15} \Rightarrow_{r10} d$

$\mathcal{A}'_{19} : \lceil b \in \overline{a} \rceil$ $\quad$ $\mathcal{A}''_3 : \lceil d \in \overline{\lceil c \in \overline{\lceil b_1 \Rightarrow_{r2} b \rceil} \rceil} \rceil$

The grounded extension of the framework from $\mathcal{PAT}_\beta$, and β's beliefs are as follows (again, with only object-level arguments considered for clarity):

$$E_G(\mathcal{PAT}_\beta) = \left\{ \begin{array}{lll} \mathcal{A}_3 : a_1, & \mathcal{A}_4 : a_2, & \mathcal{A}_5 : b_1, \\ \mathcal{A}_{14} : \mathcal{A}_5 \Rightarrow_{r2} b, & \mathcal{A}_{22} : d_1, & \mathcal{A}_{23} : d_2, \\ \mathcal{A}_{24} : \mathcal{A}_{14}, \mathcal{A}_{15} \Rightarrow_{r10} d & & \end{array} \right\}$$

$B^c_\beta = \{a_1, a_2, b_1, b, d_1, d_2, d\}$

The abstract framework derivable from $\mathcal{PAT}_\beta$ is shown in Figure 4.2, with islands removed for clarity.

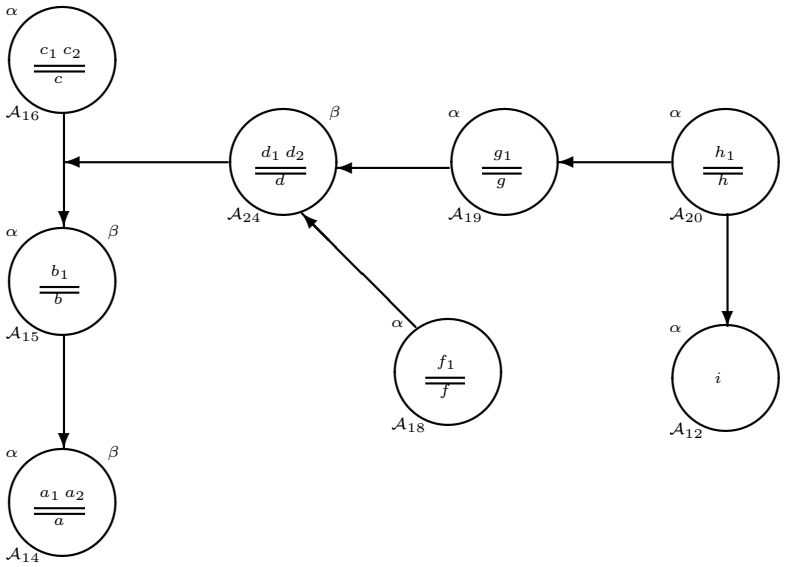

Figure 4.3: Combined framework from $\mathcal{PAT}_\alpha$ and $\mathcal{PAT}_\beta$

The combined framework from $\mathcal{PAT}_\alpha$ and $\mathcal{PAT}_\beta$ can be seen in Figure 4.3. The acceptability of arguments is not labelled, because acceptability is participant-specific, however arguments have been labelled to show its source (PAT_α, $\mathcal{PAT}_\beta$ or both). Neither participant will have knowledge of this combined framework, merely their own part of it and, following the dialogue, the parts communicated by their opponent. However, it is provided here to show that, if the dialogue were exhaustive (w.r.t. each agent's beliefs), α could win a dialogue with a topic of $\tau = a$, if $\mathcal{A}_{18}$ were not private. However, since it is private, β would win the dialogue.

Consider the dialogue fragment in Table 4.1. For clarity and brevity, only new commitments are explicitly represented at each move.

id	pl	loc	t	C_{pl}
1	α	$claim(\{a\})$	$-$	$\{a\}$
2	β	$why(\{a\})$	1	$\emptyset$
3	α	$claim(\{a_1, a_2, \lceil a_1, a_2 \Rightarrow a \rceil\})$	2	$C^1_\alpha \cup \{a_1, a_2, \lceil a_1, a_2 \Rightarrow a \rceil\}$
4	β	$claim(\{b, \lceil b \in \overline{a_1} \rceil\})$	3	$\{b, \lceil b \in \overline{a_1} \rceil\}$
5	α	$why(\{b\})$	4	C^3_α
6	β	$claim(\{b_1, \lceil b_1 \Rightarrow b \rceil\})$	5	$C^4_\beta \cup \{b_1, \lceil b_1 \Rightarrow b \rceil\}$
7	α	$claim(\{c, \lceil c \in \overline{\lceil b_1 \Rightarrow b \rceil} \rceil\})$	6	$C^5_\alpha \cup \{c, \lceil c \in \overline{\lceil b_1 \Rightarrow b \rceil} \rceil\}$
8	β	$claim(\{d, \lceil d \in \overline{\lceil c \in \overline{\lceil b_1 \Rightarrow b \rceil} \rceil} \rceil\})$	7	$C^6_\beta \cup \{d, \lceil d \in \overline{\lceil c \in \overline{\lceil b_1 \Rightarrow b \rceil} \rceil} \rceil\}$
9	α	$why(\{d\})$	8	C^7_α
10	β	$claim(\{d_1, \lceil d_1 \Rightarrow d \rceil\})$	9	$C^8_\beta \cup \{d_1, \lceil d_1 \Rightarrow d \rceil\}$

Table 4.1: Dialogue fragment

At this point, α's original claim of a is not defended, because while α defeated β's counter-claim of b (with c), the undercut from c on $b_1 \Rightarrow b$ was itself defeated by β's claim (and justification) of d. α does possess an acceptable defeater of d (f, on d_1), however f_1 is marked as a private belief, and thus both it and f are incommunicable by α. This leaves α with no option but to concede, then retract their original claim of a. This in turn will require a retraction of at least one element of the support for a (a_1, a_2 and/or $\lceil a_1, a_2 \Rightarrow a \rceil$), along with (if α so wishes) a justification of the retraction.

However, for now, the example shall be left in its current state (i.e. without α conceding or retracting anything). It will be continued in Chapter 6, where it will be demonstrated that argument revision is an effective tool in either determining what to retract, or deciding whether or not it is better to be dishonest in order to defend the original claim.

4.7 Summary

In this chapter, a simple dialogue framework, SPD, was specified. The framework is based on the ASPIC$^+$ framework for argumentation, and its associated meta- extensions defined in Chapter 3.

Before specifying the framework, each participant's beliefs were established. Each participant in an SPD dialogue possesses a personal argumentation theory, whose abstract

framework is evaluated under grounded semantics; the conclusions of all arguments in the grounded extension (closed under strict rule application) are what constitute beliefs.

The notion of private arguments was also introduced. Private arguments are acceptable arguments that, using some presently unspecified criteria, the participant does not want to reveal. Using this set of private arguments, and the beliefs determined as above, we established a set of communicable beliefs for each participant — those beliefs they are prepared to share in a dialogue.

In terms of the dialogue framework itself, SPD differs from previous ASPIC$^+$ dialogue frameworks [Weide and Dignum, 2011] in that its locutions are defined over formulae of the logical language $\mathcal{L}$, as opposed to arguments. This allows for a repeated claim-challenge-concede process instead of relying on entire arguments being communicated at once. This influences the construction of a shared argumentation theory, which keeps a deductively monotonic record of the dialogue, and is used to determine what previous claims are no longer acceptable.

Finally, a running example was presented. This example demonstrated the dialogue protocol up until a point where one of the participants would appear to have to concede then retract. However, the example has been left at this point and will be picked up again in Chapter 6.

Chapter 5

Argument Revision

5.1 Introduction

In this chapter, *Argument Revision* is formally defined and specified in terms of a system of structured argumentation. While in classic models of belief revision, the process of modifying a belief set is guided by a qualitative, arbitrary entrenchment ordering over the beliefs, the model presented here instead examines quantifiable effects of a revision on the system in arriving at a measured and justifiable determination of minimal change.

Before considering the process of Argument Revision, consideration must first be given to what the goals of the process are. In belief revision, the aim is to update a knowledge base, either by adding new beliefs (both with and without consideration for consistency), or by removing existing ones. While it might seem obvious to directly translate these concepts to Argument Revision (i.e. to add new arguments or remove existing ones), to do so would overlook an important principle in the system of argumentation being used — that of argument acceptability.

Recall from Chapter 4 that a participant in a dialogue based on SPD uses the conclusions of arguments in the grounded extension of (the abstract framework derived from) its personal argumentation theory as a basis for its beliefs. If a certain argument were to disappear from the grounded extension, the participant would no longer believe its

conclusion (assuming no other acceptable arguments for that same conclusion appear in the grounded extension). Alternatively, if a new argument were added to the theory, and that argument were acceptable, the participant would believe its conclusion.

An argument can disappear from an extension in one of two ways — either it is defeated by another (acceptable) argument, or it is removed from the argumentation theory completely. Similarly, an argument can be added to an extension either through adding it to the theory, and there being no acceptable defeaters of it, or through removing or making unacceptable all acceptable defeaters.

Argument Revision is, therefore, more flexible than belief revision, because instead of modifying a flat belief set, we are instead modifying the underlying model (i.e. system of argumentation) from which belief is derived. Since the derivation of belief depends on both the existence and acceptability of arguments, Argument Revision provides two possible ways in which to add and remove beliefs (structure and acceptability) whereas flat belief sets provide only one (structure alone).

The goal of an Argument Revision process is one of:

1. ensuring that a certain argument is removed from, or made unacceptable in, an argumentation theory; or

2. ensuring that a certain argument is added to, and/or made acceptable in, an argumentation theory.

To reach either of these goals, the process of Argument Revision involves modifying an argumentation theory and its constituent argumentation systems. Arguments are constructed through applying rules to a knowledge base, which means the removal of arguments from an argumentation theory is achieved through modifying either the knowledge base or set of rules. Acceptability of arguments is computed based on defeat relations between them, which in turn is arrived at based on 1) contrariness between formula; 2) the preference ordering over arguments (which in itself is partly derived from preferences over the knowledge base and defeasible rules); and 3) strictness and firmness properties of

arguments, based on which type of rule(s) (i.e. strict or defeasible) and types of premise (i.e. axiom, ordinary or assumption) the arguments use.

It is not, however, necessary to specify different ways of performing each type of modification detailed above. Recall that in meta-argumentation (Chapter 3), every object-level component (i.e. rule, contrary and preference) there exists a corresponding meta-level argument which concludes the meta-level representation of that component. As with object-level arguments, meta-level arguments are also open to revision and so, for instance, a rebutting attack on argument $\mathcal{B}$ by argument $\mathcal{A}$ can be removed by removing the meta-argument $\mathcal{A}'$ with $Conc(\mathcal{A}') = \lceil Conc(\mathcal{A}) \in \overline{Conc(\mathcal{B})} \rceil$. Alternatively, $\mathcal{A}'$ could be made unacceptable by introducing an argument that defeats it or by modifying meta-meta-arguments to introduce an attack, modify preferences etc.

The key observation here is that all forms of Argument Revision can be performed by either adding or removing (meta-) arguments, which involves modifying a knowledge base and/or rule sets. Since rules are themselves represented by meta-arguments, we can specify only a single method of Argument Revision: knowledge base modification.

In this chapter, we provide a formal account of Argument Revision in terms of the ASPIC$^+$ framework, incorporating definitions of Argument Revision and expansion, functions for modifying an argumentation system's knowledge base and a determination of minimal change based on quantifiable, measurable effects of a revision. Properties of three special types of revision (rule-, contrary- and preference-based) are also explored.

This chapter proceeds as follows: in section 5.2 the principles of argument contraction and expansion are defined and explained; in section 5.3 the process of Argument Revision is explained; in section 5.4, rule, preference and contrariness-based Argument Revision is explained; in section 5.5 the properties of Argument Revision are explored and in section 5.6 measures of minimal change and their use in determining a minimal change are defined.

5.2 Argument contraction and expansion

Argument Revision is sub-divided into two broad processes — *argument contraction* and *argument expansion*. In both processes, the goal relates to the *acceptability* of arguments and not necessarily (in the case of contraction at least) their presence in the resultant argumentation theory.

The goal of argument contraction is to ensure certain arguments are no longer acceptable, while expansion ensures that they are. As is the case with belief revision, we assume it is not possible to uniquely specify Argument Revision functions, because there may be multiple possible methods of achieving the goal [Gärdenfors, 1988]. This will be further explained in the definitions of the operators.

In subsequent sections and chapters, the notation $\mathcal{A} \in E(\mathcal{AT})$ means there is an acceptable argument in the abstract framework derived from the meta-argumentation theory $\mathcal{AT}$, under some unspecified, unique-extension semantics[1], subsumed by complete semantics.

5.2.1 Principles of Argument Revision

Before formally defining operators for Argument Revision, it is first necessary to specify the exact nature of Argument Revision, in terms of what a revision process aims to achieve.

In the classic AGM theory of belief revision, there are three broad processes — *contraction*, where a belief is removed from a belief set, *expansion*, where new beliefs are added to a belief set with no consideration for consistency, and *revision* where new beliefs are added and (if necessary) existing beliefs are modified to bring about consistency [Alchourrón et al., 1985].

In order to translate these concepts into those suitable for Argument Revision, defined in terms of the ASPIC$^+$ framework, we must consider certain features of the framework, and its intended use. Firstly, as noted in section 5.1, when using systems of argumentation

[1]The restriction to unique-extension semantics is imposed because translations between multi-extension semantics and belief are unclear, and it is beyond the scope of this thesis to explore them.

built on Dung's [1995] abstract theory (as the ASPIC$^+$ framework [Prakken, 2010] does), the acceptability of arguments provides a convenient method of establishing beliefs. Thus, in order to revise beliefs, it is possible to modify only the acceptability of arguments, without necessarily modifying the arguments themselves. Note that this does not preclude the modification of arguments, but it does provide an additional method of revision, which may result in fewer changes (see section 5.6 for further exploration of minimal change).

Secondly, consistency in the ASPIC$^+$ framework is different from that found in the belief sets used in the AGM theory. Prakken [2010] proves that the rationality postulates of Caminada [2007a] hold for his version of the ASPIC$^+$ framework, provided certain conditions are met. These rationality postulates include two relating to consistency:

- **Direct consistency** — the set of conclusions of all arguments in an extension is consistent
- **Indirect consistency** — the closure of the set of conclusions of all arguments in an extension under strict-rule application is consistent

The conditions required for these postulates to hold are that (i) the preference ordering over arguments is reasonable (Chapter 3, p.41, Definition 3.2.5); and (ii) that the argumentation theory is well-formed (Chapter 3, p.40). In the present work, we shall always use either the last-link or weakest-link principles for establishing a preference ordering, both of which Prakken [2010] proves to be reasonable, and thus that condition will always be satisfied.

Well-formedness, however, depends entirely on the composition of the argumentation theory and does not have a pre-defined algorithm that guarantees it for all theories. During an Argument Revision process, it is entirely possible that a theory is yielded which is not well-formed — consider in the simplest case adding to the knowledge base a formula that is a contrary of the consequent of a strict rule.

Thus, a second overall principle in Argument Revision is that the resultant argumentation theory be well-formed. This ensures that the set of conclusions in an extension

remains consistent, and thus the agent's beliefs do likewise.

5.2.2 Argument contraction

Argument contraction is the process of modifying an argumentation theory to yield a new, well-formed argumentation theory such that certain arguments that are acceptable in the original theory are not acceptable in the new theory. As previously noted, multiple methods of performing a contraction may exist; while each of these methods will have yielded an argumentation theory that satisfies the goal (i.e. the specified arguments are not acceptable in it), each of these theories will be different, due to having had different changes made to it.

Given an argumentation theory $\mathcal{AT}$ and a set of arguments $\mathcal{S}$, $\mathcal{AT}\dot{-}\mathcal{S}$ is a maximal (w.r.t. set inclusion) set of argumentation theories where $\forall \mathcal{AT}^{-} \in \mathcal{AT}\dot{-}\mathcal{S}$:

- $\mathcal{AT}^{-}$ is well-formed
- $\forall \mathcal{A} \in \mathcal{S}, \mathcal{A} \notin E(\mathcal{AT}^{-})$

We refer to $\mathcal{AT}^{-}$ as a *contraction* of $\mathcal{AT}$ by $\mathcal{S}$.

Different methods will be used to reach each argumentation theory in $\mathcal{AT}\dot{-}\mathcal{S}$; these possible methods will be described in further detail in section 5.3.

5.2.3 Argument expansion

Argument expansion is the process of modifying an argumentation theory into a new, well-formed argumentation theory, such that a given set of arguments are both present, and acceptable in the new theory. Again, it is not possible to uniquely specify an expansion function due to the possible different methods of achieving it, so as with argument contraction, the output is not a distinct argumentation theory, but a set of all theories that satisfy the properties.

Given an argumentation theory $\mathcal{AT}$ and a set of arguments $\mathcal{S}$, $\mathcal{AT}\dot{+}\mathcal{S}$ is a maximal (w.r.t. set inclusion) set of argumentation theories where $\forall \mathcal{AT}^{+} \in \mathcal{AT}\dot{+}\mathcal{S}$:

- $\mathcal{AT}^+$ is well-formed
- $\forall \mathcal{A} \in \mathcal{S}, \mathcal{A} \in E(\mathcal{AT}^+)$

We refer to $\mathcal{AT}^+$ as an *expansion* of $\mathcal{AT}$ by $\mathcal{S}$.

5.3 Process of Argument Revision

The revision of an argumentation theory, whether it be through contraction or expansion, is carried out by modifying one or more of the argumentation systems in the theory, which in turn results in arguments being added, removed, or changing acceptability. For example, a contraction may be carried out by adding a new argument that makes one or more of the arguments provided to the contraction operator unacceptable; similarly, an expansion may be carried out by removing an argument that defeats one of the arguments provided to the expansion operator.

The process of revising an argumentation theory is recursive. For each input argument to an expansion or contraction, the goal is achieved through modifying arguments that interact with it, whether they be sub-arguments, attacking arguments or meta-level arguments for rules, preference and contrariness.

The complete removal and addition of arguments can be achieved through modifying the knowledge base and/or rules in an argumentation system, while acceptability can also be altered, by modifying the preference and contrariness relations. The use of meta-argumentation allows each of these modifications to be specified by a single process, where the knowledge base in an argumentation system is modified.

Recall that for every rule and preference or contrariness in an argumentation system, there is a corresponding (possibly atomic) meta-argument for it in a higher-level system. Thus, instead of modifying rules, preferences and contrariness directly, we can instead perform an Argument Revision process with respect to their corresponding meta-level arguments. And, since meta-argumentation systems are argumentation systems, a single

method of knowledge base modification can be specified, and applied to any knowledge base, in any argumentation system in an argumentation theory.

Even small modifications to the knowledge based in an argumentation system can have a big impact on the arguments in the argumentation theory. For every argument, its premises are a subset of the knowledge base, so removing only one of those premises will result in the entire argument (and a subset of its sub-arguments) being lost. Similarly, adding only one formula to the knowledge base can allow multiple new arguments to be constructed, if all other required premises are already present.

Where the intention is to change the acceptability of arguments, the process is less straightforward than simply adding or removing arguments, because changing acceptability can be achieved in several different ways:

- Adding a new argument (through adding one or more formulae to the knowledge base) which acts as a (possibly indirect) defeater or defender of the original argument
- Removing an argument (through removing one or more formulae from the knowledge base) which acts as a (possibly indirect[2]) defender or defeater of the original argument.
- Revising the meta-arguments for contrariness and preference relations, which involve the original argument.

The knowledge base in an argumentation system can be modified by both adding and removing formulae. The result of an addition is a new argumentation system, whose argumentation theory is possibly not well-formed. Before formally defining the concepts of formula addition and removal, a definition of equality between argumentation systems is first provided. This definition allows certain properties of expansion and removal to be expressed. For clarity, the contrariness function, $^-$, is represented by cf.

[2]Dung [1995] defines indirect defence as: $\mathcal{A}$ indirectly defends $\mathcal{B}$ iff there exists an even-length path (with non-zero length) from $\mathcal{A}$ to $\mathcal{B}$ in the argument graph

Definition 5.3.1 *(Argumentation system equality)*

Two argumentation systems $\mathcal{AS}_1$ and $\mathcal{AS}_2$ are equal iff:

- $\mathcal{L}(\mathcal{AS}_1) = \mathcal{L}(\mathcal{AS}_2)$
- $cf(\mathcal{AS}_1) = cf(\mathcal{AS}_2)$
- $\mathcal{R}(\mathcal{AS}_1) = \mathcal{R}(\mathcal{AS}_2)$
- $\mathcal{K}(\mathcal{AS}_1) = \mathcal{K}(\mathcal{AS}_2)$
- *the partial pre-orders on $\mathcal{K}(\mathcal{AS}_1)$ and $\mathcal{R}(\mathcal{AS}_1)$ are, respectively, identical to those on $\mathcal{K}(\mathcal{AS}_2)$ and $\mathcal{R}(\mathcal{AS}_2)$*

In other words, two (object-level) argumentation systems are equal iff all their components (language, contrariness, rules, preferences and knowledge base) are, respectively, identical.

5.3.1 Formula removal

When a formula is removed from a knowledge base, the result is a new argumentation system that is identical to the original argumentation system, except for the knowledge base which lacks the removed formula.

Formally, this is captured by the *formula removal function*, which takes as input an argumentation system and some formula, and provides as output the new argumentation system whose knowledge base does not contain that formula:

Definition 5.3.2 ***Formula removal function***

$\mathcal{AS} - \phi$*:* $\mathcal{K}(\mathcal{AS} - \phi) = \mathcal{K}(\mathcal{AS}) \backslash \{\phi\}$

To show that an argumentation theory has had a formula removed from the knowledge base of one of its argumentation systems, the notation $\mathcal{AT} - (\phi, \mathcal{AS}_n)$ is used. For instance, if the formula ϕ is removed from the knowledge base of $\mathcal{AS}_1$, then:

$$\mathcal{AT} - (\phi, \mathcal{AS}_1) = \langle \{\mathcal{AS}_1 - \phi, \ldots, \mathcal{AS}_n\}, \{\mathcal{K}_1 \backslash \{\phi\}, \ldots, \mathcal{K}_n\}, \preceq \rangle$$

Certain properties of the formula removal function can be demonstrated. Firstly, if the input formula is not present in the knowledge base of $\mathcal{AS}$ (that is, $\phi \notin \mathcal{K}(\mathcal{AS})$), the output argumentation system is identical to the output:

Proposition 5.3.1 *If $\phi \notin \mathcal{K}(\mathcal{AS})$, $\mathcal{AS} - \phi = \mathcal{AS}$*

Proof: if $\phi \notin \mathcal{K}(\mathcal{AS})$ then $\mathcal{K}(\mathcal{AS}) \backslash \{\phi\} = \mathcal{K}(\mathcal{AS})$. Thus $\mathcal{K}(\mathcal{AS} - \phi) = \mathcal{K}(\mathcal{AS})$ and hence $\mathcal{AS} - \phi = \mathcal{AS}$. □

It is possible to remove multiple formulae at once. This is the same as performing a sequence of removals, where the input argumentation system for the n^{th} removal is the output from the $(n-1)^{th}$ removal.

Proposition 5.3.2 $\mathcal{AS} - (\phi_1 \wedge \ldots \wedge \phi_n) = \mathcal{AS} - \phi_1 - \ldots - \phi_n$

Proof: Consider first the knowledge base of $\mathcal{AS} - (\phi_1 \wedge \ldots \wedge \phi_n)$: $\mathcal{K}(\mathcal{AS} - \phi_1 \wedge \ldots \wedge \phi_n) = \mathcal{K}(\mathcal{AS}) \backslash \{\phi_1, \ldots, \phi_n\} = \mathcal{K}(\mathcal{AS}) \backslash \{\phi_1\} \backslash \ldots \backslash \{\phi_n\}$,

Consider now the knowledge base of $\mathcal{AS} - \phi_1 - \ldots - \phi_n$: $\mathcal{K}(\mathcal{AS} - \phi_1 - \ldots - \phi_n) = \mathcal{K}(\mathcal{AS} - \phi_1 - \ldots - \phi_{n-1}) \backslash \{\phi_n\} = \mathcal{K}(\mathcal{AS} - \phi_1 - \ldots - \phi_{n-2}) \backslash \{\phi_{n-1}\} \backslash \{\phi_n\} = \ldots = \mathcal{K}(\mathcal{AS}) \backslash \{\phi_1\} \backslash \ldots \backslash \{\phi_n\} = \mathcal{K}(\mathcal{AS} - \phi_1 \wedge \ldots \wedge \phi_n)$ □

Formula removal also always preserves well-formedness.

Proposition 5.3.3 *If $\mathcal{AT}$ is well-formed, then $\mathcal{AT} - (\phi, \mathcal{AS})$ is also well-formed.*

Proof: Since $\mathcal{R}(\mathcal{AS}) = \mathcal{R}(\mathcal{AS} - \phi)$, we consider only the knowledge base in $\mathcal{AS} - \phi$. If $\mathcal{AT}$ is well-formed, then $\mathcal{K}(\mathcal{AS})$ satisfies [Prakken, 2010, Definition 6.8]. Since $\mathcal{K}(\mathcal{AS} - \phi) \subseteq \mathcal{K}(\mathcal{AS})$, $\mathcal{K}(\mathcal{AS} - \phi)$ also satisfies the definition. □

5.3.2 Formula addition

The *formula addition function* governs the addition of a formula to the knowledge base of an argumentation system.

One issue of adding information to the knowledge base is knowing which subset it should be placed in. The nature of the formula addition function is that it adds formulae simply to introduce new arguments, with no justification (where justification could, for instance, come from another agent). We therefore classify added formulae as assumptions.

Definition 5.3.3 ***Formula addition function***

$\mathcal{AS} + \phi$: $\mathcal{K}_a(\mathcal{AS} + \phi) = \mathcal{K}_a(\mathcal{AS}) \cup \{\phi\}$

Remark 9 *Note also that because* $\mathcal{K}_a(\mathcal{AS}) \subseteq \mathcal{K}(\mathcal{AS})$, $\mathcal{K}(\mathcal{AS} + \phi) = \mathcal{K}(\mathcal{AS}) \cup \{\phi\}$

Similar to the formula removal function, to show that an argumentation theory has had a formula added to the knowledge base of one of its argumentation systems, the notation $\mathcal{AT} + (\phi, \mathcal{AS}_n)$ is used. For instance, if the formula ϕ is added to the knowledge base of $\mathcal{AS}_1$, then:

$$\mathcal{AT} + (\phi, \mathcal{AS}_1) = \langle \{\mathcal{AS}_1 + \phi, \ldots, \mathcal{AS}_n\}, \{\mathcal{K}_1 \cup \{\phi\}, \ldots, \mathcal{K}_n\}, \preceq \rangle$$

The formula addition function possesses several properties that assist in an Argument Revision process. Firstly, the function is deductively monotonic (that is, no arguments, acceptable or otherwise, are lost when expanding an argumentation system).

Proposition 5.3.4 *The formula addition function,* $+$*, is deductively monotonic.*

Proof: $\forall \mathcal{A} \in \mathcal{A}rgs(\mathcal{AS})$, $Prem(\mathcal{A}) \subseteq \mathcal{K}(\mathcal{AS})$. *For some* $\phi \in \mathcal{L}$, $\mathcal{K}(\mathcal{AS}) \subseteq \mathcal{K}(\mathcal{AS} + \phi)$. *Thus,* $\forall \mathcal{A} \in \mathcal{A}rgs(\mathcal{AS})$, $Prem(\mathcal{A}) \subseteq \mathcal{K}(\mathcal{AS} + \phi)$. □

Furthermore, there is no change to an argumentation system if the input formula to an expansion is already an assumption in the knowledge base, or if the input formula to a removal is not in the knowledge base:

Proposition 5.3.5 *If for* $\mathcal{AS} + \phi$, $\phi \in \mathcal{K}_a(\mathcal{AS})$, $\mathcal{AS} + \phi = \mathcal{AS}$.

Proof: $\mathcal{K}_a(\mathcal{AS} + \phi) = \mathcal{K}_a(\mathcal{AS}) \cup \{\phi\}$. *Since* $\phi \in \mathcal{K}_a(\mathcal{AS})$, $K_a(\mathcal{AS} + \phi) = \mathcal{K}_a(\mathcal{AS})$. □

Proposition 5.3.6 *If for* $\mathcal{AS} - \phi$, $\phi \notin \mathcal{K}(\mathcal{AS})$, $\mathcal{AS} - \phi = \mathcal{AS}$.

Proof: $\mathcal{K}(\mathcal{AS} - \phi) = \mathcal{K}(\mathcal{AS}) \backslash \{\phi\}$. *Since* $\phi \notin \mathcal{K}(\mathcal{AS})$, $K(\mathcal{AS} - \phi) = \mathcal{K}(\mathcal{AS})$. □

Both the formula removal and formula addition functions yield new argumentation systems and, hence, a new argumentation theory. However, it's possible that further modifications are required to these new systems — for instance, if adding a formula brings about a new argumentation theory that is not well-formed. As with the initial removal or expansion, there may be multiple possible modifications, which again raises the question of which to choose.

In order to model the possible choices when performing an Argument Revision process, we use a structure called a change graph.

5.3.3 Change graphs

When revising an argumentation theory, through expansion or contraction, there may be multiple possible ways in which the goal can be achieved (that is, arriving at an argumentation theory that satisfies the properties of whichever revision process is being used). Furthermore, some methods have more than one stage — for instance, if adding a formula brings about an argumentation theory that is not well-formed — with each stage also presenting a choice as to the next possible stage.

This raises first the question of how to identify all possible methods of expanding or contracting an argumentation theory; then, there is the question of exactly which method should be chosen. In this section, the issue of identifying the possible methods will be addressed; choosing between them is then addressed in section 5.6.

Recall that the revision of an argumentation theory is performed through modifying the knowledge bases in the argumentation systems within the theory, with modifications continuing to take place until the properties of the chosen revision (i.e. expansion or contraction) are satisfied. One such sequence of modifications can be seen as a route, or path, from the original argumentation theory to an argumentation theory that has been expanded or contracted with respect to the input arguments.

Thus, to model the possible ways in which an argumentation theory can be either expanded or contracted (again, with respect to a set of input arguments) by modifying

the knowledge bases of its argumentation systems, we define a structure called a *change graph*. A change graph can be used for either expansion or contraction and its definition is not dependant on either. For convenience, we use the notation Π' to represent a set of argumentation theories $\{\mathcal{AT}^{\pm}_1, \ldots, \mathcal{AT}^{\pm}_n\}$ that are revisions of $\mathcal{AT}$, but do not necessarily satisfy the criteria for expansion or contraction.

Definition 5.3.4 *A change graph $CG(\mathcal{AT}, \Pi')$ for the revisions of $\mathcal{AT}$ to a set of argumentation theories $\Pi' = \{\mathcal{AT}^{\pm}_1, \ldots, \mathcal{AT}^{\pm}_n\}$ is a directed acyclic graph $\langle \Upsilon, \Omega \rangle$ where:*

- $\Upsilon \subseteq \Pi$ *(where Π is the set of all possible argumentation theories)*
- $\Pi' \subseteq \Upsilon$ *is minimal (w.r.t. set inclusion) in that $\forall \mathcal{AT}' \in \Pi'$, $\neg\exists \mathcal{AT}'' \in \Pi'$, $\mathcal{K}(\mathcal{AT})\Delta\mathcal{K}(\mathcal{AT}'') \subset \mathcal{K}(\mathcal{AT})\Delta\mathcal{K}(\mathcal{AT}')$*
- ***(Atomic change)*** $\Omega \subseteq \Upsilon \times \Upsilon$ *where $\forall \omega \in \Omega$ such that $\omega = (\mathcal{AT}', \mathcal{AT}'')$, we have that $\mid \mathcal{K}(\mathcal{AT}')\Delta\mathcal{K}(\mathcal{AT}'') \mid = \pm 1$*

The minimality constraint on Π' in Definition 5.3.4 is a pruning exercise that discounts any argumentation theories arrived at by making unnecessary additions and/or removals, by considering the set of net changes between $\mathcal{AT}$ and some $\mathcal{AT}' \in \Pi'$: there can exist no argumentation theory $\mathcal{AT}'' \in \Pi'$ where the set of net changes in $\mathcal{AT}''$ is a proper subset of the set of net changes in $\mathcal{AT}'$. The atomic change condition of Definition 5.3.4 mandates that on each edge of a change graph, only one removal or one addition in one knowledge base is performed.

If any $\mathcal{AT}' \in \Pi'$ satisfies the properties of $\mathcal{AT}\dot{-}\mathcal{S}$ (it is well-formed and no arguments in $\mathcal{S}$ are acceptable in $\mathcal{AF}_{\mathcal{AT}'}$), then $\mathcal{AT}'$ is a contraction of $\mathcal{AT}$ by $\mathcal{S}$ and hence $\mathcal{AT}' \in \mathcal{AT}\dot{-}\mathcal{S}$. Similarly, if any $\mathcal{AT}' \in \Pi'$ satisfies the properties of $\mathcal{AT}\dot{+}\mathcal{S}$ (it is well-formed and all arguments in $\mathcal{S}$ are acceptable in $\mathcal{AF}_{\mathcal{AT}'}$) then $\mathcal{AT}'$ is an expansion of $\mathcal{AT}$ by $\mathcal{S}$ and hence $\mathcal{AT}' \in \mathcal{AT}\dot{+}\mathcal{S}$.

5.4 Rule, preference and contrariness-based Argument Revision

This chapter has thus far focused on modifying the knowledge base in an argumentation system in order to revise an argumentation theory. It is also possible to achieve the goal of a revision process by modifying the rules, preferences and contrariness in an argumentation system (possibly in conjunction with knowledge base modifications). It is in making these modifications that meta-argumentation is used.

Recall the definition of a meta-argumentation system (Chapter 3, p.44, Definition 3.3.1), in which every component of an object-level system is the conclusion of an argument in the associated meta-level system. This argument may be atomic (i.e. the meta-level representation of the component is an element of the knowledge base in the meta-system), however it could equally be a conclusion derivable from a set of antecedents and an inference rule. In either case, it is still possible to model the addition or removal of the component using the same change graph, because each edge represents the modification of a knowledge base in any of the argumentation systems within an argumentation theory.

Revising arguments for contrariness and preferences involves the same process as revising object-level arguments — modifying the knowledge base in the meta-argumentation system with respect to the argument(s) for the contrariness or preference. Rule-based revision, however, is more complex, because there are several ways in which changes to rules can bring about changes to the arguments and their acceptability in an argumentation theory.

5.4.1 Rule-based Argument Revision

Revising the rules in an argumentation system can be used to either introduce new arguments, or modify existing ones such that they are removed altogether, or have their strength modified in order to change the success or otherwise of an attack.

Rule addition and removal

Using rules to introduce new or removing existing arguments is similar to the overall Argument Revision process, in that in order to add or remove rules at the object-level, we modify the arguments for them at the meta-level. Modifying these arguments is an Argument Revision process in itself, carried out in the same way is any other.

Rule reclassification

The second method of rule-based revision does not involve adding or removing rules, but instead reclassifying them. The set of rules in an argumentation system is divided into those that are strict (i.e. hold without exception) and those that are defeasible (i.e. generally hold, but may not under some exceptional circumstances).

To reclassify a rule is to change a strict rule to defeasible, or a defeasible rule to strict. Reclassification of rules does not impact on the construction of arguments, but it does impact on their properties — consider, in the simplest case, reclassifying a rule used by a strict argument. Since that rule would become defeasible, the argument too would be defeasible and, hence, weaker.

As a more concrete example, consider the following knowledge base and rule sets in an argumentation system $\mathcal{AS}_i$:

- $\mathcal{K}_p = \{p, q\}$
- $\mathcal{R}_s = \{(p \rightarrow s)\}$
- $\mathcal{R}_d = \{(q \Rightarrow \neg s)\}$

The following arguments can be constructed:

- $\mathcal{A}_1 : q$
- $\mathcal{A}_2 : p$
- $\mathcal{A}_3 : \mathcal{A}_2 \rightarrow s$

- $\mathcal{A}_4 : \mathcal{A}_1 \Rightarrow \neg s$

If $\mathcal{R}_s$ were then closed under transposition (Chapter 3, p.41, Definition 3.2.7), $Cl_{trs}(\mathcal{R}_s) = \{p \rightarrow s, \neg s \rightarrow \neg p\}$. This allows $\mathcal{A}_4$ to be extended with $\mathcal{A}^+ : \mathcal{A}_4 \rightarrow \neg p$, which in turn defeats $\mathcal{A}_2$ and hence also $\mathcal{A}_3$. The resultant abstract framework, evaluated under grounded semantics, can be seen in Figure 5.1. Arguments with a solid border are "in" and arguments with a dashed border are "undecided".

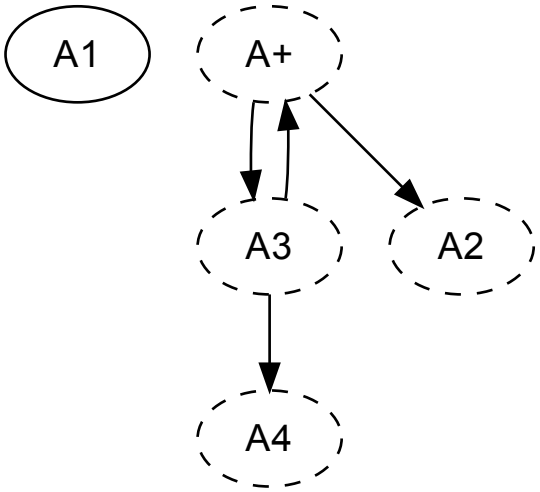

Figure 5.1: Framework using strict rule $p \rightarrow s$ with $\mathcal{R}_s$ closed under transposition

If, however, we were to modify $\mathcal{AS}_i$, into $\mathcal{AS}'_i$ such that:

- $\mathcal{K}'_p = \mathcal{K}_p$
- $\mathcal{R}'_s = \emptyset$ (removal of $p \rightarrow s$)
- $\mathcal{R}'_d = \{(p \Rightarrow s), (q \Rightarrow \neg s)\}$ (addition of $p \Rightarrow s$)

the defeat relations change. The resultant abstract framework, again evaluated under grounded semantics with the same labelling convention, can be seen in Figure 5.2. There are two changes to the framework; firstly, instead of a strict argument $\mathcal{A}_3$ for s, there is now

a defeasible argument $\mathcal{A}'_3$; while this still defeats $\mathcal{A}_4$, it is now reciprocated. Furthermore, and of more significance, $\mathcal{A}^+$ is no longer an argument and thus $\mathcal{A}_2$ is not defeated and so is now labelled "in".

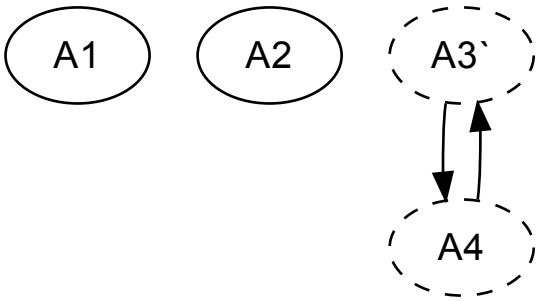

Figure 5.2: Framework using defeasible rule $p \Rightarrow s$

The same example can be used to illustrate the opposite process — a defeasible rule changing to strict. If our initial argumentation system contained the rule $p \Rightarrow s$, we would start with the framework in Figure 5.2; if, in a new system, this rule were then substituted for $p \rightarrow s$, we would obtain the framework in Figure 5.1.

Rule reclassification is characterised through the use of meta-argumentation. Recall the characterisation of strict and defeasible rules introduced in definition 3.3.2. A rule is strict iff its meta-level representation has no contraries; otherwise, it is defeasible. Thus, a strict rule at the object-level can be re-classified into a defeasible rule if we were to introduce a contrary to it at the meta-level; similarly, a defeasible rule at the object-level can be re-classified into a strict rule if we were to remove all contraries to it at the meta-level.

In the above example of strict-to-defeasible reclassification, we could introduce a contrary $\phi \in \overline{\lceil p \Rightarrow s \rceil}$ in the argumentation system $\mathcal{AS}_{i+1}$. Note that is it not necessary for an argument for ϕ to be present in the meta-level system (acceptable or otherwise); what is important is the presence of the contrariness relation. The opposite process (i.e. defeasible to strict reclassification) would take place through the removal of the existing

contrary or contraries that caused the rule to be currently classified as defeasible.

Reclassifying rules is similar to premise-based Argument Revision, in that one small change can have a significant impact on the argumentation theory and the resultant abstract framework, as the simple example presented above shows. In the case of strict-to-defeasible reclassification, the example has shown the complete removal of one argument and another made acceptable; conversely, defeasible-to strict-reclassification results in one argument being added and another changing from acceptable to undecided.

Introducing a contrary in an argumentation system requires a corresponding meta-level argument for it. Therefore in order to introduce a meta-level contrariness relation with respect to an object-level rule, we must introduce an argument for it at the next meta-level — that is, an argumentation system at level $i + 2$, given that the object level is level i and its meta-level is level $i + 1$.

The process of introducing an argument at any level is identical — perform an argument expansion process with respect to the argument that is to be introduced, which in the case of arguments that are currently not present will require the addition of knowledge base formulae, rules or both.

5.4.2 Contrariness and preference-based Argument Revision

Contrariness and preferences are used in order to determine defeat between arguments, with the former being used to determine the initial attack and the latter determining whether or not it is successful (i.e. results in defeat). Using contrariness and preferences in Argument Revision is thus a method of changing the acceptability of arguments, similar to rule reclassification. However, the actual process of modifying contrariness and preferences is more analogous to rule addition and removal, in that it is carried out through modifying the meta-level arguments for the appropriate components.

To illustrate the effect of contrariness-based Argument Revision, consider the argumentation framework in Figure 5.3, where in the underlying argumentation theory ($\mathcal{AT}$) $Conc(\mathcal{A}_1) \in \overline{Conc(\mathcal{A}_3)}$, $Conc(\mathcal{A}_2) \in \overline{Conc(\mathcal{A}_1)}$ and $\mathcal{A}_3 \in Sub(\mathcal{A}_2)$ (which, in turn,

have corresponding meta-arguments).

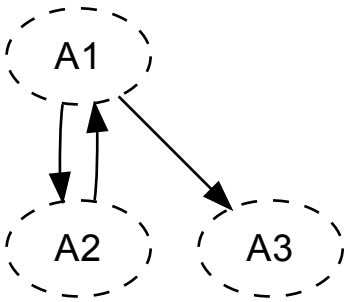

Figure 5.3: Argumentation framework with all arguments "undecided"

If $\mathcal{AT}$ were contracted with respect to the meta-argument for contrariness between $\mathcal{A}_1$ and $\mathcal{A}_3$ (i.e. $\mathcal{AT}\dot{-}\{\mathcal{A}'_1\}$, where $\mathcal{A}'_1 : \ulcorner Conc(\mathcal{A}_1) \in \overline{Conc(\mathcal{A}_3)} \urcorner$), the argumentation framework would change, as shown in Figure 5.4.

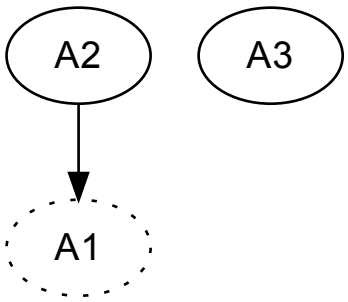

Figure 5.4: Argumentation framework with $\mathcal{A}_2$ and $\mathcal{A}_3$ "in" and $\mathcal{A}_1$ "out"

5.5 Properties of Argument Revision

In sections 5.3.1 and 5.3.2, several properties of the formula removal and addition operators were described. In this section, the properties of combining these operators will be examined, along with an exploration of the operators in terms of the properties of change in abstract argumentation specified by Cayrol et al. [2010].

5.5.1 Combining removal and addition

When a removal and an addition are combined, the order in which the operators are applied is important, especially if the first operator does not cause any changes to the argumentation system. For instance, a removal can only be "undone" (through addition) if the input formula was an assumption in the original knowledge base. This shows the same behaviour as the recovery postulate in AGM [Gärdenfors, 1988].

Proposition 5.5.1 $(\mathcal{AS} - \phi) + \phi = \mathcal{AS}$ *iff* $\phi \in \mathcal{K}_a(\mathcal{AS})$

Proof: From basic set theory:

- *If* $\phi \notin \mathcal{K}_a(\mathcal{AS})$, $\mathcal{AS} - \phi = \mathcal{AS}$. *However,* $\phi \in \mathcal{K}_a((\mathcal{AS} - \phi) + \phi)$, *hence* $(\mathcal{AS} - \phi) + \phi \neq \mathcal{AS}$.
- *If* $\phi \in \mathcal{K}_a(\mathcal{AS})$, $\mathcal{K}(\mathcal{AS} - \phi) = \mathcal{K}_a(\mathcal{AS})\backslash\{\phi\}$. $\mathcal{K}_a((\mathcal{AS} - \phi) + \phi) = \mathcal{K}_a(\mathcal{AS} - \phi) \cup \{\phi\} = (\mathcal{K}_a(\mathcal{AS})\backslash\{\phi\}) \cup \{\phi\} = \mathcal{K}_a(\mathcal{AS})$

□

Similarly, an expansion can only be "undone" (through removal) only if the input formula was not already in the original knowledge base.

Proposition 5.5.2 $\mathcal{AS} + \phi - \phi = \mathcal{AS}$ *iff* $\phi \notin \mathcal{K}(\mathcal{AS})$

Proof: From basic set theory:

- *If* $\phi \in \mathcal{K}_a(\mathcal{AS})$, $\mathcal{AS} + \phi = \mathcal{AS}$. *However,* $\mathcal{AS} - \phi \neq \mathcal{AS}$.

- *If* $\phi \notin \mathcal{K}(\mathcal{AS})$, $\mathcal{K}(\mathcal{AS}+\phi) = \mathcal{K}(\mathcal{AS}) \cup \{\phi\}$. $\mathcal{K}((\mathcal{AS}+\phi)-\phi) = \mathcal{K}(\mathcal{AS}+\phi) \backslash \{\phi\} = (\mathcal{K}(\mathcal{AS}) \cup \{\phi\}) \backslash \{\phi\} = \mathcal{K}(\mathcal{AS})$

□

5.5.2 Structural properties

The properties of change in abstract argumentation described by Cayrol et al. [2010] can equally be applied to ASPIC$^+$ structured argumentation, because it instantiates Dung's abstract approach. The structural properties provided in Cayrol et al.'s work describe the effects on the set of extensions when adding and removing arguments and interactions (attacks). These are summarised in Table 5.1.

Property for a change operation	Characterisation of the property
the change is **decisive**	$E = \emptyset$ or $E = \{\{\}\}$ or $\mid E \mid > 2$ and $\mid E' \mid = 1$ and $E' \neq \{\{\}\}$
the change is **restrictive**	$\mid E \mid > \mid E' \mid > 2$
the change is **questioning**	$\mid E \mid < \mid E' \mid$
the change is **destructive**	$E \neq \emptyset$ and $E \neq \{\{\}\}$ $E' = \emptyset$ or $E' = \{\{\}\}$
the change is **expansive**	$\mid E \mid = \mid E' \mid$ and $\forall \varepsilon_i^j \in E'$, $\exists \varepsilon_i \in E, \varepsilon_i \subset \varepsilon_i^j$
the change is **conservative**	$E = E'$
the change is **altering**	$\mid E \mid = \mid E' \mid$ and $\exists \varepsilon_i \in E$ s.t. $\forall \varepsilon_j^i \in E'$, $\varepsilon_i \not\subseteq \varepsilon_j^i$

Table 5.1: Structural properties for a change operation, from [Cayrol et al., 2010]

An Argument Revision process will constitute exactly one of the above changes. There are some limitations, however; expansion can never be *destructive*, while *conservative* can only hold of expansion and contraction under certain specific circumstances.

Proposition 5.5.3 *An expansion can never be destructive.*

Proof: For some $\mathcal{AT}$ *and set of arguments* $\mathcal{S} = \{\mathcal{A}_1, \ldots, \mathcal{A}_n\}$, $S \subseteq E(\mathcal{AT}\dot{+}\mathcal{S})$. *However, if* $\mathcal{AT}\dot{+}\mathcal{S}$ *were destructive,* $E(\mathcal{AT}\dot{+}\mathcal{S}) = \emptyset$ *or* $E(\mathcal{AT}\dot{+}\mathcal{S}) = \{\{\}\}$ □

Proposition 5.5.4 *A contraction* $\mathcal{AT}\dot{-}\mathcal{S}$ *is conservative if* $\forall \mathcal{A} \in \mathcal{S}$, $\mathcal{A} \notin E(\mathcal{AT})$.

Proof: If $\forall \mathcal{A} \in \mathcal{S}$, $\mathcal{A} \notin E(\mathcal{AT})$, $\mathcal{AT}\dot{-}\mathcal{S} = \mathcal{AT}$ *and hence* $E(\mathcal{AT}) = E(\mathcal{AT}\dot{-}\mathcal{S})$. □

Proposition 5.5.5 *An expansion $\mathcal{AT}\dot{+}\mathcal{S}$ is conservative if $\forall \mathcal{A} \in \mathcal{S}$, $\mathcal{A} \in E(\mathcal{AT})$.*

Proof: If $\forall \mathcal{A} \in \mathcal{S}$, $\mathcal{A} \notin E(\mathcal{AT})$, $\mathcal{AT}\dot{+}\mathcal{S} = \mathcal{AT}$ and hence $E(\mathcal{AT}) = E(\mathcal{AT}\dot{+}\mathcal{S})$. □

In section 5.6, these properties are further explored in terms of measures of minimal change.

5.6 Measures of minimal change

One of the main principles in the AGM theory of belief revision is that of *minimal change* — when beliefs are revised, the process is carried out with as small an impact as possible on the beliefs that remain. This is not measured solely in terms of logical consequences, but also through an entrenchment ordering placed on beliefs, where entrenchment is a measure of how important a belief is; the higher the degree of entrenchment, the more important the belief. Conversely, the lower the degree of entrenchment, the more likely the agent is to give up that belief during a belief revision process.

A major drawback of entrenchment orderings is the criteria used to determine them. Gärdenfors [1988] assumes that only qualitative criteria are suitable, because they solve the problem of uniquely specifying a revision (or expansion) function. However, this is a rather substantial trade-off, because the price for allowing a unique revision (expansion) function is to introduce some unspecified, and possibly arbitrary criteria for evaluating beliefs. Furthermore, in a dynamic environment an agent will possibly, by definition, receive new information of which it was not previously aware; the use of qualitative methods makes it almost impossible for the agent to autonomously evaluate the importance of this information and, thus, appropriately place it in an entrenchment ordering.

If we are to look beyond the concept of an entrenchment ordering as a way of determining minimal change when revising an ASPIC$^+$ argumentation theory, the features of the framework itself provide certain clues. An obvious feature to include is the structure of the arguments themselves; an argumentation theory is revised through modifying the knowledge bases in its argumentation systems in order to add or remove arguments. However,

any element in a knowledge base may be a premise in more than one argument, resulting in not only an input argument being eliminated from the system, and the theory, but also other arguments that are completely unrelated. Consider the knowledge base and rules in an argumentation system in the argumentation theory $\mathcal{AT}$:

- $\mathcal{K}_p = \{a\}$
- $\mathcal{R}_d = \{a \Rightarrow b, a \Rightarrow c\}$

With arguments:

- $\mathcal{A}_1$: a
- $\mathcal{A}_2$: $\mathcal{A}_1 \Rightarrow b$
- $\mathcal{A}_3$: $\mathcal{A}_1 \Rightarrow c$

Assume that we have no further information (i.e. relating to contrariness, preferences etc.) and wish to contract $\mathcal{AT}$ with respect to $\mathcal{A}_2$; that is, $\mathcal{AT}\dot{-}\{\mathcal{A}_2\}$. The only possible method is to either remove a from the knowledge base, or $a \Rightarrow b$ from the rules (through revising its meta-level argument). Removing a would result in not only $\mathcal{A}_1$ being removed from the theory, but also $\mathcal{A}_3$ and therefore has a greater effect than removing $a \Rightarrow b$.

A further consideration relates to the acceptability of arguments. If a previously acceptable argument becomes unacceptable following a revision process, any arguments of which it was the sole defeater would become acceptable; conversely, an unacceptable argument that becomes acceptable would render unacceptable any arguments it is a defeater of.

To expand the above example, consider the following extensions to the knowledge base and rules, and new contrariness:

- $\mathcal{K}_p = \{a, d\}$
- $\mathcal{R}_d = \{a \Rightarrow b, a \Rightarrow c, d \Rightarrow e\}$

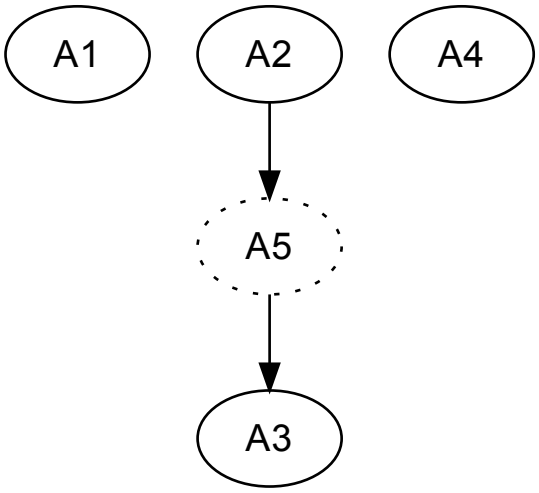

Figure 5.5: Argumentation framework with $\mathcal{A}_5$ "out"

- $b \in \overline{e}$, $e \in \overline{c}$

This yields the arguments:

- $\mathcal{A}_1$: a
- $\mathcal{A}_2$: $\mathcal{A}_1 \Rightarrow b$
- $\mathcal{A}_3$: $\mathcal{A}_1 \Rightarrow c$
- $\mathcal{A}_4$: d
- $\mathcal{A}_5$: $\mathcal{A}_4 \Rightarrow e$

The abstract framework derived from $\mathcal{AT}$, evaluated under grounded semantics (where a solid border means the argument is "in" and a dotted border means the argument is "out"), is shown in Figure 5.5.

Again, if we were to contract $\mathcal{AT}$ with respect to $\mathcal{A}_2$, and assuming it were carried out through removing the argument completely, the framework would change such that $\mathcal{A}_5$ becomes acceptable and $\mathcal{A}_3$ becomes unacceptable. This is shown in Figure 5.6.

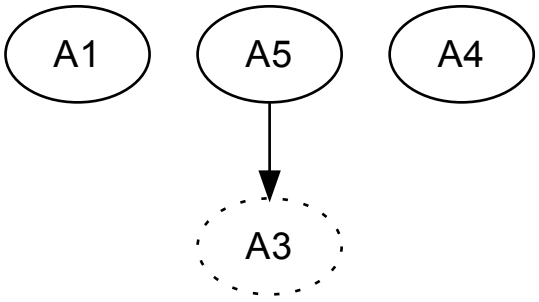

Figure 5.6: Argumentation framework with $\mathcal{A}_5$ "in"

Thus, we have four measurable effects of a revision process on an argumentation theory:

- Argument loss — when removing elements of an argumentation system's knowledge base, at the very least the atomic arguments constructed from those elements will be lost to the system and the theory. If those arguments are then sub-arguments in other arguments, those other arguments will also be lost.

- Acceptability loss — when making either type of modification to an argumentation system's knowledge base, arguments that defend other arguments may be lost, which in turn makes those arguments unacceptable. Alternatively, a contrariness (attack) or preference relation may be added or removed, resulting in new or removed defeat.

- Argument gain — when adding elements to an argumentation system's knowledge base, those new elements will, at the very least, result in new atomic arguments being created in the system and the theory. When then considered with the rules in the system, they may also become sub-arguments in other, larger arguments.

- Acceptability gain — when making either type of modification to an argumentation system's knowledge base, arguments that are the sole defeaters of other arguments may be lost or made unacceptable, which in turn makes those defeated arguments

acceptable. Alternatively, a contrary (attack) or preference relation may be added or removed, resulting in defeaters being defeated, or the attack that results in a defeat being lost.

To capture these four effects, each is defined in terms of a function related to formula additions and removals. The definition of each function involves an argumentation system $\mathcal{AS}$ in an argumentation theory $\mathcal{AT}$, and a formula ϕ in the knowledge base in $\mathcal{AS}$. Υ_G is used to represent the set of all argumentation theories in a change graph, and the type of change (addition or removal) unspecified; however, in subsequent chapters and sections we will, where necessary, make explicit the type of change with a superscript $+$ (for addition) or $-$ (for removal). Where it is necessary to represent a change that is left unspecified, we again use $\pm$.

The first function, the *argument drop function* identifies those acceptable arguments that are completely lost when removing or adding a formula.

Definition 5.6.1 *The* argument drop function Δ_A *for an extension* E *(under some single-extension semantics subsumed by complete semantics) in an argumentation theory* $\mathcal{AT}$*:*
Δ_A*:* $\Pi \times \Pi \rightarrow 2^{\mathcal{A}rgs(\mathcal{AT})}$,
$\Delta_A(\mathcal{AT}, \mathcal{AT}^{\pm}) = \{\mathcal{A} \mid \mathcal{A} \in E(\mathcal{AT}), \mathcal{A} \notin \mathcal{A}rgs(\mathcal{AT}^{\pm})\}$

The output of the argument drop function is always an empty set when adding formulae.

Proposition 5.6.1 $\forall \phi \in \mathcal{L}_n$ *with* $\mathcal{AT}^{\pm} = \mathcal{AT} + (\phi_1, \mathcal{AS}) + \ldots + (\phi_n, \mathcal{AS})$ *then* $\Delta_A(\mathcal{AT}, \mathcal{AT}^{\pm}) = \emptyset$

Proof: Consider the opposite: $\Delta_A(\mathcal{AT}, \mathcal{AT}^{\pm}) \neq \emptyset$*. Thus,* $\forall \mathcal{A} \in \Delta_A(\mathcal{AT}, \mathcal{AT}^{\pm})$*,* $\mathcal{A} \in \mathcal{A}rgs(\mathcal{AT})$ *and* $\mathcal{A} \notin \mathcal{A}rgs(\mathcal{AT}^{\pm})$*, and hence* $Prem(\mathcal{A}) \subseteq \mathcal{K}(\mathcal{AT})$ *and* $Prem(\mathcal{A}) \not\subseteq \mathcal{K}(\mathcal{AT}^{\pm})$*. However, from proposition 5.3.4,* $\mathcal{K}(\mathcal{AT}) \subseteq \mathcal{K}(\mathcal{AT} + (\phi, \mathcal{AS}))$ *and hence* $\mathcal{K}(\mathcal{AT}) \subseteq \mathcal{K}(\mathcal{AT} + (\phi_1, \mathcal{AS}) + \ldots + (\phi_n, \mathcal{AS}))$*. Contradiction!* □

The second function, the *acceptability drop function*, identifies those arguments that lose acceptability when removing or adding a formula.

Definition 5.6.2 *The* acceptability drop function Δ_S *for an extension* E *(under some single-extension semantics subsumed by complete semantics) in an argumentation theory* $\mathcal{AT}$*:*

Δ_S*:* $\Pi \times \Pi \rightarrow 2^{\mathcal{A}rgs(\mathcal{AT})}$,

$\Delta_S(\mathcal{AT}, \mathcal{AT}^{\pm}) = \{\mathcal{A} \in E(\mathcal{AT}), \mathcal{A} \in \mathcal{A}rgs(\mathcal{AT}^{\pm}), \mathcal{A} \notin E(\mathcal{AT}^{\pm})\}$

The output of the acceptability drop function is always a conflict-free set when adding a single formula.

Proposition 5.6.2 $\forall \phi \in \mathcal{L}$, $\Delta_S(\mathcal{AT}, \mathcal{AT} \pm (\phi, \mathcal{AS}))$ *is conflict-free in* $\mathcal{AF}_{\mathcal{AT}}$

Proof: Consider an extension E *under some single-extension semantics subsumed by complete semantics. From definition 5.6.2,* $\mathcal{A} \in \Delta_S(\mathcal{AT}, \mathcal{AT} \pm (\phi, \mathcal{AS}))$ *iff* $\mathcal{A} \in E(AT)$ *and* $\mathcal{A} \notin E(\mathcal{AT} \pm (\phi, \mathcal{AS}))$*. Thus* $\Delta_S(\mathcal{AT}, \mathcal{AT} \pm (\phi, \mathcal{AS})) \subseteq E(\mathcal{AT})$*. Since* $E(\mathcal{AT})$ *is conflict-free [Dung, 1995], all subsets of* $E(\mathcal{AT})$ *are also conflict-free.* □

The third function, the *argument gain function*, identifies those arguments that the system gains when removing or adding a formula.

Definition 5.6.3 *The* argument gain function Γ_A*:*

Γ_A*:* $\Pi \times \Pi \rightarrow 2^{\mathcal{A}rgs(\mathcal{AT}^{\pm})}$,

$\Gamma_A(\mathcal{AT}, \mathcal{AT}^{\pm}) = \{\mathcal{A} \mid \mathcal{A} \notin \mathcal{A}rgs(\mathcal{AT}), \mathcal{A} \in \mathcal{A}rgs(\mathcal{AT}^{\pm})\}$

The argument gain function always yields an empty set when removing formulae.

Proposition 5.6.3 $\forall \phi \in \mathcal{L}_n$ *with* $\mathcal{AT}^{\pm} = \mathcal{AT} - (\phi_1, \mathcal{AS}) - \ldots - (\phi_n, \mathcal{AS})$ *then*

$\Gamma_A(\mathcal{AT}, \mathcal{AT}^{\pm}) = \emptyset$

Proof: Consider the opposite: $\mathcal{A} \notin \mathcal{A}rgs(\mathcal{AT})$, $\mathcal{A} \in \mathcal{A}rgs(\mathcal{AT}^{\pm})$*. Thus,* $Prem(\mathcal{A}) \not\subseteq \mathcal{K}(\mathcal{AT})$ *and* $Prem(\mathcal{A}) \subseteq \mathcal{K}(\mathcal{AT}^{\pm})$*. However,* $\mathcal{K}(\mathcal{AT}^{\pm}) = \mathcal{K}(\mathcal{AT}) \backslash \{\phi_1, \ldots, \phi_n\}$*, hence* $Prem(\mathcal{A}) \subseteq \mathcal{K}(\mathcal{AT})$ *and* $\mathcal{A} \in \mathcal{A}rgs(\mathcal{AT})$*. Contradiction!* □

The final function, the *acceptability gain function*, identifies those arguments that were acceptable in the input argumentation system and have remained in the output system, but have lost acceptability.

Definition 5.6.4 *The* acceptability gain function Γ_S*:*

Γ_S*:* $\Pi \times \Pi \rightarrow 2^{\mathcal{A}rgs(\mathcal{AT})}$,

$\Gamma_S(\mathcal{AT}, \mathcal{AT}^{\pm}) = \{\mathcal{A} \mid \mathcal{A} \notin E(\mathcal{AT}), \mathcal{A} \in E(\mathcal{AT}^{\pm})\}$

The drop and gain functions for each type of change (i.e. argument and acceptability) are linked to each other, in that no arguments (resp. acceptability) that are dropped are also gained, and vice versa.

Proposition 5.6.4 *For* $X \in \{A, S\}$, $\Delta_X(\mathcal{AT}, \mathcal{AT}^{\pm}) \cap \Gamma_X(\mathcal{AT}, \mathcal{AT}^{\pm}) = \emptyset$

For $X \in \{A, S\}$, $\Delta_X(\mathcal{AT}, \mathcal{AT}^{\pm}) \cap \Gamma_X(\mathcal{AT}, \mathcal{AT}^{\pm}) = \emptyset$

Proof: Consider the opposite in terms of argument drop and gain: $\mathcal{A} \in \Delta_A(\mathcal{AT}, \mathcal{AT}^{\pm})$, $\mathcal{A} \in \Gamma_A(\mathcal{AT}, \mathcal{AT}^{\pm})$. *From definitions 5.6.1 and 5.6.3,* $\mathcal{A} \notin \mathcal{A}rgs(\mathcal{AT}^{\pm})$ *and* $\mathcal{A} \in \mathcal{A}rgs(\mathcal{AT}^{\pm})$ *respectively. Contradiction!*

Consider the opposite in terms of acceptability drop and gain: $\mathcal{A} \in \Delta_S(\mathcal{AT}, \mathcal{AT}^{\pm})$, $\mathcal{A} \in \Gamma_S(\mathcal{AT}, \mathcal{AT}^{\pm})$. *From definitions 5.6.2 and 5.6.4,* $\mathcal{A} \notin E(\mathcal{AT}^{\pm})$, $\mathcal{A} \in E(\mathcal{AT}^{\pm})$. *Contradiction!* □

Given these properties, it can be proven that any argumentation theory can always contracted by any subset of the arguments in the theory.

Proposition 5.6.5 *For any* $\mathcal{S} \subseteq \mathcal{A}rgs(\mathcal{AT})$, $\mathcal{AT}\dot{-}\mathcal{S}$ *is always defined.*

Proof: $\forall \mathcal{A} \in \mathcal{A}rgs(\mathcal{AT})$*, for any* $\phi \in Prem(\mathcal{A})$, $\mathcal{A} \in \Delta_A^-(\mathcal{AT}, \mathcal{AT} - (\phi, \mathcal{AS}))$. *Thus, in the extreme case, given* $\Phi = \bigcup_{\mathcal{A} \in \mathcal{S}} Prem(\mathcal{A})$, $\mathcal{S} \subseteq \bigcup_{\phi \in \Phi} \Delta_A^-(\mathcal{AT}, \mathcal{AT} - (\phi, \mathcal{AS}))$ □

However, the same does not hold unconditionally for expansion. Recall that an expansion must (i) have all input arguments acceptable in the derived framework and (ii) lead to a well-formed argumentation theory. However, simple examples can show how these principles are violated.

Consider arguments $\mathcal{A}_1 : p$ and $\mathcal{A}_2 : \neg p$. It is not possible for both $\mathcal{A}_1$ and $\mathcal{A}_2$ to be acceptable under a unique extension semantics in an argumentation theory containing the arguments, and thus $\mathcal{AT}\dot{+}\{\mathcal{A}_1, \mathcal{A}_2\}$ is undefined. Similarly, consider $\mathcal{B}_1 : p \rightarrow q$

and $\mathcal{B}_2 : r \rightarrow \neg q$; an argumentation theory containing these arguments would not be well-formed, since the consequences of two or more strict rules cannot be in conflict with one another. Thus, for $\mathcal{AT}\dot{+}\mathcal{S}$ to be defined, $\mathcal{S}$ must be conflict-free.

Recall that when an argumentation theory is expanded, existing arguments may need contracted to ensure the input arguments are all acceptable. This allows us to take advantage of proposition 5.6.5 in proving that, provided $\mathcal{S}$ is conflict-free, $\mathcal{AT}\dot{+}\mathcal{S}$ is always defined.

Proposition 5.6.6 *For any conflict-free set of arguments $\mathcal{S}$ (based on $\mathcal{L}$), $\mathcal{AT}\dot{+}\mathcal{S}$ is always defined.*

Proof: Since for any $\mathcal{S}' \subseteq \mathcal{A}rgs(\mathcal{AT})$ $\mathcal{AT}\dot{-}\mathcal{S}'$ is always defined (through, in the extreme case, all arguments in $\mathcal{S}$ being completely removed), it follows that $\mathcal{A}rgs(\mathcal{AT}\dot{-}\mathcal{A}rgs(\mathcal{AT})) = \emptyset$. Thus if $\mathcal{AT}\dot{+}\mathcal{S}$ incorporates the contraction by $\mathcal{A}rgs(\mathcal{AT})$, there remain no arguments in conflict with $\mathcal{S}$ and thus $\mathcal{AT}\dot{+}\mathcal{S}$ is defined. □

Finally, in order to use a change graph to determine minimal change, we first, assign costs to edges in a change graph, based on outputs of the four functions for measuring minimal change. We do this by computing the *path cost* for each path from $\mathcal{AT}$ to every argumentation theory in Π'.

Definition 5.6.5 *Given a change graph $CG(\mathcal{AT}, \Pi') = \langle \Upsilon, \Omega \rangle$, the* path cost *from $\mathcal{AT}$ to $\mathcal{AT}' \in \Pi'$ is*

$V(\mathcal{AT}, \mathcal{AT}') = \mid \Delta_A(\mathcal{AT}, \mathcal{AT}') \cup \Delta_S(\mathcal{AT}, \mathcal{AT}') \cup \Gamma_A(\mathcal{AT}, \mathcal{AT}') \cup \Gamma_S(\mathcal{AT}, \mathcal{AT}') \mid$

Comparing the output of V for the paths from $\mathcal{AT}$ to each argumentation theory in Π' allows the minimal change(s) to be identified.

5.7 Change graph example

Having now defined a structure for determining minimal change, a simple example based only on object-level arguments will now be presented to illustrate the concepts. A fuller

example, including meta-level arguments, will be given in Chapter 6 as a continuation of the running example introduced in Chapter 4.

Consider an argumentation theory $\mathcal{AT}$ with a single argumentation system $\mathcal{AS}$, which in turn contains the following knowledge base and rules:

- $\mathcal{K}_p = \{a_1, a_2, b_1, b_2\}$
- $\mathcal{R}_d = \{(a_1, a_2 \Rightarrow a), (a_1 \Rightarrow x), (a_2 \Rightarrow y), (b_1, b_2 \Rightarrow b)\}$

and that $a_2 \in \overline{b_1}$ and $b \in \overline{c}$.

The arguments in the theory are as follows:

- $\mathcal{A}_1 : a_1$
- $\mathcal{A}_2 : a_2$
- $\mathcal{A}_3 : b_1$
- $\mathcal{A}_4 : b_2$
- $\mathcal{A}_5 : c_1$
- $\mathcal{A}_6 : \mathcal{A}_1, \mathcal{A}_2 \Rightarrow a$
- $\mathcal{A}_7 : \mathcal{A}_1 \Rightarrow x$
- $\mathcal{A}_8 : \mathcal{A}_2 \Rightarrow y$
- $\mathcal{A}_9 : \mathcal{A}_3, \mathcal{A}_4 \Rightarrow b$
- $\mathcal{A}_{10} : \mathcal{A}_5 \Rightarrow c$

The abstract framework derived from $\mathcal{AT}$, and evaluated under grounded semantics, is shown in Figure 5.7. For clarity, arguments with no interactions (attacking or attacked) are not rendered in the diagram, but all are acceptable.

The grounded extension is $\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_4, \mathcal{A}_6, \mathcal{A}_7, \mathcal{A}_8, \mathcal{A}_{10}\}$.

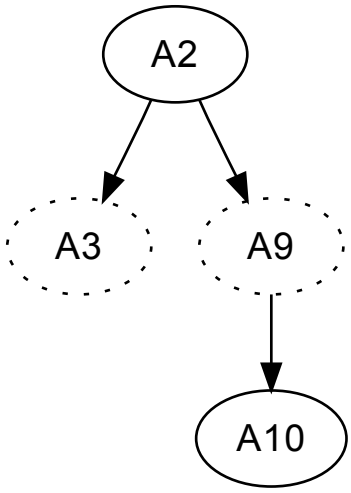

Figure 5.7: Abstract framework from $\mathcal{AT}$

Assume that we wish to achieve $\mathcal{AT}\dot{-}\{\mathcal{A}_6, \mathcal{A}_{10}\}$ using only premise-based Argument Revision. First, we consider $\mathcal{A}_5$; this argument contains two premises, a_1 and a_2; the effects of removing these, in terms of argument and acceptability drop and gain, are shown in Table 5.7.

ϕ	Δ_A^-	Δ_S^-	Γ_A^-	Γ_S^-
a_1	$\{\mathcal{A}_1, \mathcal{A}_6, \mathcal{A}_7\}$	$\{\}$	$\{\}$	$\{\}$
a_2	$\{\mathcal{A}_2, \mathcal{A}_6, \mathcal{A}_8\}$	$\{\mathcal{A}_{10}\}$	$\{\}$	$\{\mathcal{A}_3, \mathcal{A}_9\}$

Table 5.2: Possible initial removals in $\mathcal{AT}\dot{-}\{\mathcal{A}_6, \mathcal{A}_{10}\}$

Removing a_2 achieves the goal of the Argument Revision process — $\mathcal{A}_6$ has been lost completely, while $\mathcal{A}_{10}$ has been rendered unacceptable. Removing a_1 only achieves half of the goal, in that $\mathcal{A}_6$ has been lost completely, but $\mathcal{A}_{10}$ remains in the theory and is acceptable. However, that does not mean to remove a_1 is the wrong course of action — we can consider a further change; this could be to either remove $\mathcal{A}_{10}$ completely, or actively perform a change which is merely a consequence of potentially removing a_1 (in order to bring about a change of acceptability). The possible changes and their effects are captured

in Table 5.3.

ϕ	Δ_A^-	Δ_S^-	Γ_A^-	Γ_S^-
c_1	$\{\mathcal{A}_5, \mathcal{A}_{10}\}$	$\{\}$	$\{\}$	$\{\}$
a_2	$\{\mathcal{A}_2, \mathcal{A}_8\}$	$\{\mathcal{A}_{10}\}$	$\{\}$	$\{\mathcal{A}_3, \mathcal{A}_9\}$

Table 5.3: Possible initial removals in $\mathcal{AT} \dot{-} \{\mathcal{A}_6, \mathcal{A}_{10}\}$

Following the removal of a_1, removing either c_1 or a_2 will result in a contraction with respect to $\mathcal{A}_{10}$. However, removing a_2 would violate the edge-minimality condition for paths in a change graph (definition 5.3.4), because there already exists a shorter (w.r.t. edges) path from $\mathcal{AT}$ to $\mathcal{AT} \dot{-} \{\mathcal{A}_6, \mathcal{A}_{10}\}$ that sees a_2 removed. Thus we only consider the removal of c_1 after removing a_1.

Thus $\mathcal{AT} \dot{-} \{A_6, \mathcal{A}_{10}\} = \{(\mathcal{AT} - (a_1, \mathcal{AS})) - (c_1, (\mathcal{AS} - a_1)), \mathcal{AT} - (a_2, \mathcal{AS})\}$

The change graph that results from these processes is shown in Figure 5.8.

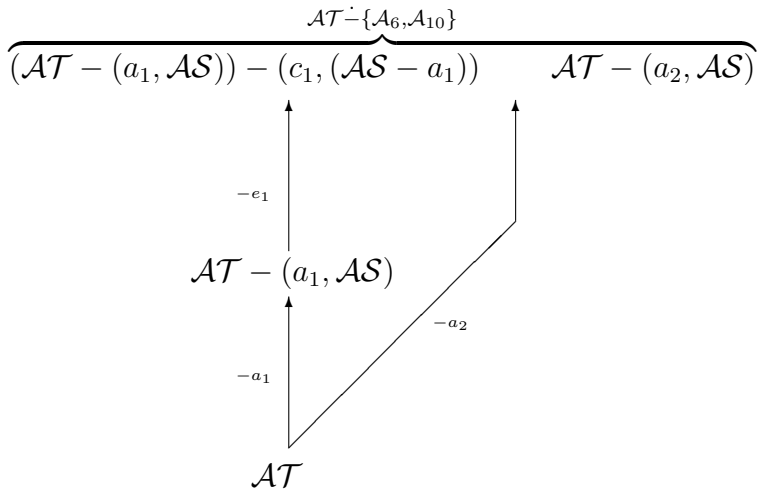

Figure 5.8: Change graph for $\mathcal{AT} \dot{-} \{\mathcal{A}_6, \mathcal{A}_{10}\}$

The path costs are as follows:

$V((\mathcal{AT}, \mathcal{AT} - (a_2, \mathcal{AS})) = \mid \{\mathcal{A}_2, \mathcal{A}_6, \mathcal{A}_8\} \cup \{\mathcal{A}_{10}\} \cup \{\mathcal{A}_3, \mathcal{A}_9\} \mid = 6$

$V(\mathcal{AT}, (\mathcal{AT} - (a_1, \mathcal{AS})) - (c_1, (\mathcal{AS} - a_1))) = \mid \{\mathcal{A}_1, \mathcal{A}_6, \mathcal{A}_7, \mathcal{A}_5, \mathcal{A}_{10}\} \mid = 5$

It can therefore be seen that removing a_1 and c_1 from the knowledge base represents the minimal change when contracting $\mathcal{AT}$ with respect to $\mathcal{A}_6$ and $\mathcal{A}_{10}$.

This example has illustrated two important principles in Argument Revision:

1. *Minimal change is not always represented by the minimal **actions***: in the example, removing a_2 achieved the goal of the Argument Revision, but resulted in greater effects on the argumentation theory than removing a_1 and c_1.

2. *Argument acceptability is an important consideration*: if argument acceptability was not considered in the example, removing a_2 would have had fewer measurable effects on the system (since a total of three arguments changed acceptability in this removal, which would have brought the total changes down to 3).

5.8 Summary

In this chapter, a model for Argument Revision in the ASPIC$^+$ framework has been presented. Argument Revision takes the form of either argument expansion, where the goal is to add and/or make acceptable a set of arguments and argument contraction, where the goal is to remove or make unacceptable a set of arguments.

Arguments can be revised either by modifying the knowledge base or, through the use of meta-argumentation, modifying the rules, preferences and contrariness information in an argumentation system. A meta-argumentation system contains arguments about the components of an object-level argumentation system, which allows these components to be revised in the same way as object-level arguments.

A structure called a change graph was defined to allow possible methods of revising an argumentation theory to be identified and reasoned about. In order to then choose which methods, four measures of minimal change were specified and used to compute path costs on a change graph. In Chapter 7, potential methods of refining the model, including additional factors in the determination of minimal change will be discussed.

Chapter 6

Argument revision in dialogue

6.1 Introduction

In this chapter, the model of argument revision presented in Chapter 5 is applied to the dialogue protocol specified in Chapter 4 to show how, in certain contexts, argument revision techniques can assist a participant in not only selecting a dialogue move, but also determining its content.

Recall that in a dialogue, a participant may have arguments it wishes to keep private (Chapter 4, p.54). Figure 6.1 incorporates private arguments into the process performed by a participant when deciding how to respond to a claim F from their opponent. If the participant agrees with the claim, they concede it; if not, they attempt to find a defeater of it. So long as the participant possesses a *non-private* defeater, they claim its conclusion; otherwise, they concede.

Following a concession, a participant may be forced by their opponent into retracting any commitments that are inconsistent with the conceded statement. While identifying and retracting those statements that make the commitment store directly inconsistent (Chapter 4, p.58, Definition 4.4.3) is relatively trivial, retracing those that make it indirectly inconsistent (Chapter 4, p.58, Definition 4.4.4) is less so. Consider, for instance, two statements p and q that, together with (a meta-level representation of) a rule $\lceil p, q \Rightarrow r \rceil$

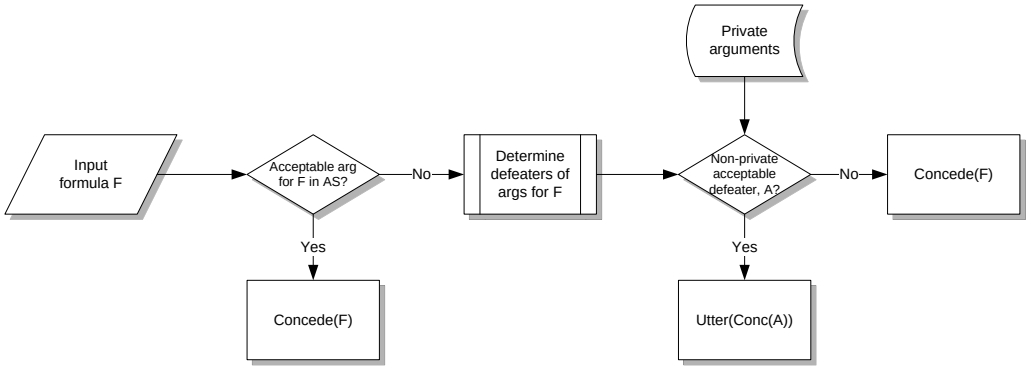

Figure 6.1: Dialogue move selection

cause inconsistency with respect to a contrary of r. p, q and $p, q \Rightarrow r$ together cause indirect inconsistency, but individually they do not. Retracting only one would solve the inconsistency, which then poses the question of which to retract.

As well as retracting, a participant does have a further option open to them — to be dishonest. This is achieved through claiming something they do not believe, in an attempt to defend their original position by defeating their opponent's counter-argument(s).

Whichever locution the participant chooses, argument revision can assist. Initially, the techniques can be used to compare retraction with dishonesty to determine which is the least costly, with respect to minimal change in both existing and potential future commitments. Then, having chosen which path to follow, the propositional content of the locution can be chosen on the same principle — i.e. by identifying what is the minimal way of justifying the retraction, or what is the minimal lie (again, both with respect to impact on existing and potential future commitments).

The chapter proceeds as follows: in sections 6.2 and 6.3, argument revision techniques are applied, respectively, to retraction of commitments and dishonesty, with the latter also providing a specification of dishonesty in terms of the ASPIC$^+$ framework and its meta-level extensions.

6.2 Commitment retraction

6.2.1 Stability adjustments

When a participant in a dialogue cannot defend a statement against a counter-argument from their opponent, they ordinarily should retract the defeated statement. However, retracting the statement alone is sometimes insufficient, because they may still hold commitment to other statements from which the retracted one is a consequence.

Recall that in an SPD dialogue, each participant's commitment store is closed (Chapter 4, p.58, Definition 4.4.2), which means that even if a statement is retracted, if there exist formulae from which that statement can be derived, it will remain in the closure. Consider, for instance, the following simple example: $C_\alpha^i = \{\phi, \psi, \lceil \phi \Rightarrow \psi \rceil\}$; retracting ψ leaves $C_\alpha^{i+1} = \{\phi, \lceil \phi \Rightarrow \psi \rceil\}$, however $Cl_C(\mathcal{C}_\alpha^{i+1}) = \{\phi, \psi, \lceil \phi \Rightarrow \psi \rceil\}$. It is, therefore, necessary to make further retractions, with or without justification, to ensure ψ can no longer be inferred.

Walton and Krabbe [1995] term the process of ensuring that a retracted statement can no longer be inferred in a commitment store a *stability adjustment*, and use the diagram in Figure 6.2 to illustrate one possible way in which a participant can achieve it (nodes represent statements and edges represent inference, with the arrows indicating the direction of support). The participant has retracted the statement P, but must also retract the circled statements to ensure that P can no longer be inferred.

A drawback of Walton and Krabbe's specification of stability adjustments is that they do not explain exactly which statements should be selected; consider Figure 6.3, which shows the same inference tree for P, but instead sees two statements on the left-hand side retracted, as opposed to the two on the right-hand side.

The same total number of retractions has taken place, so why is it that the adjustment shown in Figure 6.2 is used instead of that in Figure 6.3?

Argument revision offers a solution to this problem. By considering the retraction of commitments as an argument revision process, a participant can determine what, if any,

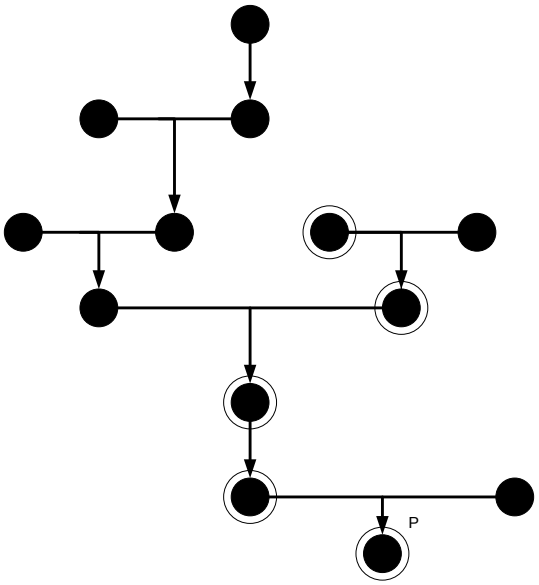

Figure 6.2: Stability adjustment from [Walton and Krabbe, 1995]

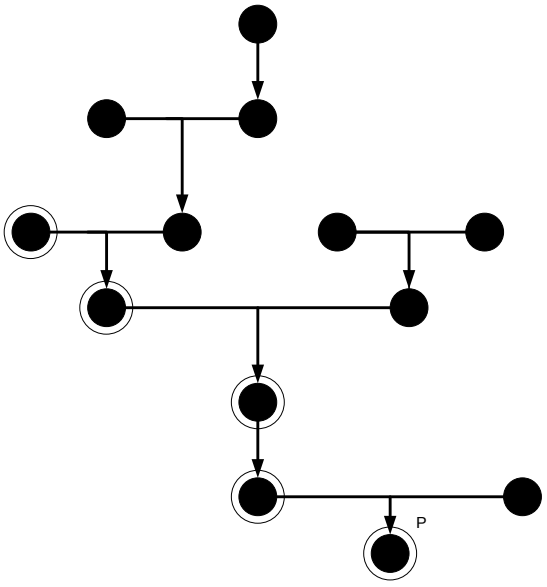

Figure 6.3: Alternative stability adjustment

methods are minimal, with respect to the measures of minimal change specified in Chapter 5.

Such a process would not, however, be carried out with respect to only existing commitments. It is also important to consider what the participant might *potentially* become committed to at a future point in the dialogue. Consider again the sample commitment store provided above. It might be that, with respect to commitments, retracting ϕ is minimal (compared to retracting $\lceil \phi \Rightarrow \psi \rceil$); but if ϕ is a premise in several as-yet unstated arguments in the participant's personal argumentation theory, those arguments would be rendered incommunicable in the remainder of the dialogue, unless the complexion changed such that ϕ was returned to the commitment store.

6.2.2 Choosing what to retract

In the remainder of this section, the example introduced in Chapter 4 will be continued, such that α chooses to concede to β and is then forced to retract commitments to resolve the resultant inconsistency.

First, we provide a recap of the arguments that α can construct in their personal argumentation theory, $\mathcal{PAT}_\alpha$:

$\mathcal{A}_1 : w_1$	$\mathcal{A}_2 : z$	$\mathcal{A}_3 : a_1$
$\mathcal{A}_4 : a_2$	$\mathcal{A}_5 : b_1$	$\mathcal{A}_6 : c_1$
$\mathcal{A}_7 : c_2$	$\mathcal{A}_8 : e_1$	$\mathcal{A}_9 : f_1$
$\mathcal{A}_{10} : g_1$	$\mathcal{A}_{11} : h_1$	$\mathcal{A}_{12} : i$
$\mathcal{A}_{13} : \mathcal{A}_1 \Rightarrow_{r8} w$	$\mathcal{A}_{14} : \mathcal{A}_3, \mathcal{A}_4 \Rightarrow_{r1} a$	$\mathcal{A}_{15} : \mathcal{A}_5 \Rightarrow_{r2} b$
$\mathcal{A}_{16} : \mathcal{A}_6, \mathcal{A}_7 \Rightarrow_{r3} c$	$\mathcal{A}_{17} : \mathcal{A}_8 \Rightarrow_{r4} e$	$\mathcal{A}_{18} : \mathcal{A}_9 \Rightarrow_{r5} f$
$\mathcal{A}_{19} : \mathcal{A}_{10} \Rightarrow_{r6} g$	$\mathcal{A}_{20} : \mathcal{A}_{11} \Rightarrow_{r7} h$	$\mathcal{A}_{21} : \mathcal{A}_4 \Rightarrow_{r9} u$
$\mathcal{A}'_1 : \lceil a_1, a_2 \Rightarrow_{r1} a \rceil$	$\mathcal{A}'_2 : \lceil b_1 \Rightarrow_{r2} b \rceil$	$\mathcal{A}'_3 : \lceil c_1, c_2 \Rightarrow_{r3} c \rceil$
$\mathcal{A}'_4 : \lceil e_1 \Rightarrow_{r4} e \rceil$	$\mathcal{A}'_5 : \lceil f_1 \Rightarrow_{r5} f \rceil$	$\mathcal{A}'_6 : \lceil g_1 \Rightarrow_{r6} g \rceil$
$\mathcal{A}'_7 : \lceil h_1 \Rightarrow_{r7} h \rceil$	$\mathcal{A}'_8 : \lceil w_1 \Rightarrow_{r8} w \rceil$	$\mathcal{A}'_9 : \lceil a_2 \Rightarrow_{r9} u \rceil$

$\mathcal{A}'_{10} : \lceil b \in \overline{a} \rceil$ $\mathcal{A}'_{11} : \lceil f \in \overline{d} \rceil$ $\mathcal{A}'_{12} : \lceil g \in \overline{d} \rceil$

$\mathcal{A}'_{13} : \lceil h \in \overline{g} \rceil$ $\mathcal{A}'_{14} : \lceil h \in \overline{i} \rceil$ $\mathcal{A}'_{15} : \lceil y \in \overline{z} \rceil$

$\mathcal{A}'_{16} : \lceil x \in \overline{w_1} \rceil$ $\mathcal{A}'_{17} : \lceil a_2 \in \overline{e_1} \rceil$ $\mathcal{A}'_{18} : \lceil x \in \overline{a_1} \rceil$

$\mathcal{A}''_1 : \lceil c \in \overline{\lceil b_1 \Rightarrow_{r2} b \rceil} \rceil$ $\mathcal{A}''_2 : \lceil y \in \overline{\lceil a_1, a_2 \Rightarrow_{r1} a \rceil} \rceil$

Table 6.1 provides a reminder of the dialogue so far.

id	pl	loc	t	C_{pl}
1	α	$claim(\{a\})$	$-$	$\{a\}$
2	β	$why(\{a\})$	1	$\emptyset$
3	α	$claim(\{a_1, a_2, \lceil a_1, a_2 \Rightarrow a \rceil\})$	2	$C^1_\alpha \cup \{a_1, a_2, \lceil a_1, a_2 \Rightarrow a \rceil\}$
4	β	$claim(\{b, \lceil b \in \overline{a_1} \rceil\})$	3	$\{b, \lceil b \in \overline{a_1} \rceil\}$
5	α	$why(\{b\})$	4	C^3_α
6	β	$claim(\{b_1, \lceil b_1 \Rightarrow b \rceil\})$	5	$C^4_\beta \cup \{b_1, \lceil b_1 \Rightarrow b \rceil\}$
7	α	$claim(\{c, \lceil c \in \overline{\lceil b_1 \Rightarrow b \rceil} \rceil\})$	6	$C^5_\alpha \cup \{c, \lceil c \in \overline{\lceil b_1 \Rightarrow b \rceil} \rceil\}$
8	β	$claim(\{d, \lceil d \in \overline{\lceil c \in \overline{\lceil b_1 \Rightarrow b \rceil} \rceil} \rceil\})$	7	$C^6_\beta \cup \{d, \lceil d \in \overline{\lceil c \in \overline{\lceil b_1 \Rightarrow b \rceil} \rceil} \rceil\}$
9	α	$why(\{d\})$	8	C^7_α
10	β	$claim(\{d_1, \lceil d_1 \Rightarrow d \rceil\})$	9	$C^8_\beta \cup \{d_1, \lceil d_1 \Rightarrow d \rceil\}$

Table 6.1: Dialogue fragment, from Chapter 4

Recall that α has been left in a situation where their original claim was not defensible. In Table 6.2, we continue the dialogue with α choosing to concede β's claim, with β in turn demanding that α resolve the resultant inconsistency in their commitment store.

id	pl	loc	t	C_{pl}
...	...	...	...	...
11	α	$concede(\{d_1, \lceil d_1 \Rightarrow d \rceil, \lceil d \in \overline{\lceil c \in \overline{\lceil b_1 \Rightarrow b \rceil} \rceil} \rceil\})$	10	$C^9_\alpha \cup \{d_1, d_1 \Rightarrow d\}$
12	β	$resolve(a)$	11	C^{10}_β

Table 6.2: Continuation of the dialogue

Having conceded d_1 and $\lceil d_1 \Rightarrow d \rceil$ to β, α's commitment store is now inconsistent because $\{d, \lceil d \in \overline{\lceil c \in \overline{\lceil b_1 \Rightarrow b \rceil} \rceil} \rceil, \lceil c \in \overline{\lceil b_1 \Rightarrow b \rceil} \rceil\} \subseteq Cl_C(C^8_\alpha)$, and their original claim of a is undefended in $\mathcal{AT}_D$.

Since α can offer no other defence of a, their only option is to retract it. However, three other formulae in the commitment store still allow a to be inferred: a_1, a_2 and $\lceil a_1, a_2 \Rightarrow a \rceil$.

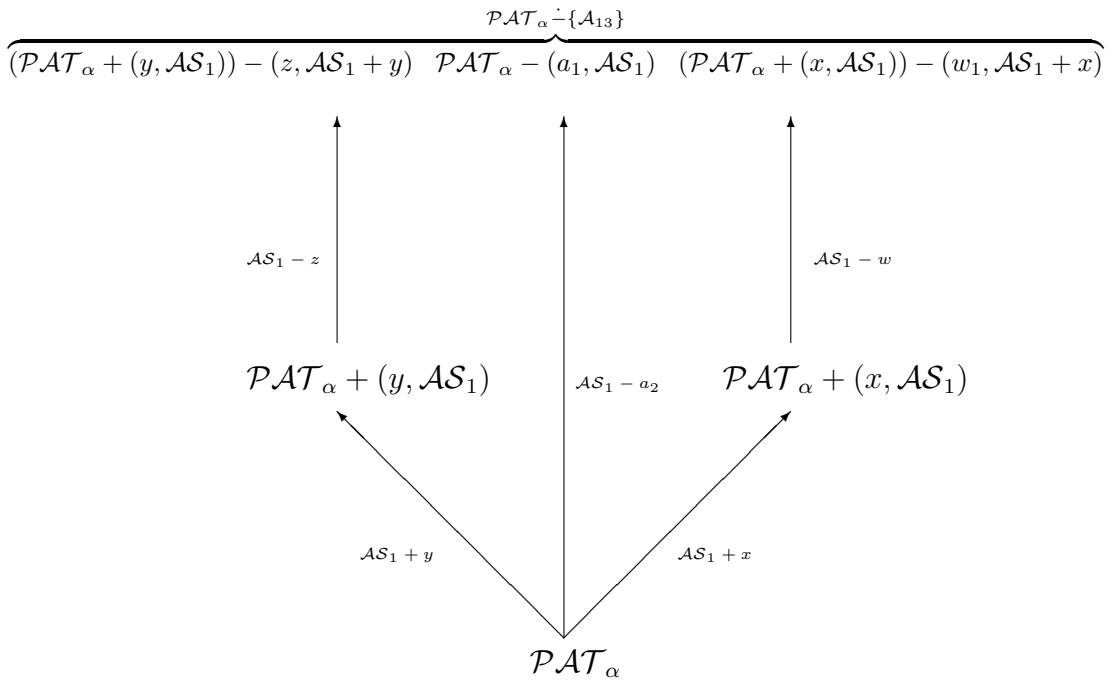

Figure 6.4: Change graph for $\mathcal{PAT}_\alpha \dot{-} \{\mathcal{A}_{13}\}$

To decide which of these to also retract, α performs an argument contraction process with respect to $\mathcal{A}_{13}$ (the argument for a). We will assume that α will, where possible, justify the retraction (Chapter 4, p.60, Definition 4.4.5). Given the arguments in α's personal argumentation theory, $\mathcal{PAT}_\alpha$, α is aware of defeaters of $\lceil a_1, a_2 \Rightarrow a \rceil$ (i.e. an undercutter for the rule) and a_1, but not a_2. Thus, α performs an argument revision processes on $\lceil a_1, a_2 \Rightarrow a \rceil$ and a_1 through introducing those defeaters (justifying), and on a_2 by simply removing it. The change graph for this is shown in Figure 6.4.

From the change graph, the outputs of the four functions for measuring minimal change (argument drop and gain and acceptability drop and gain) are shown in Table 6.3, where $\mathcal{A}_{25} : y$ and $\mathcal{A}_{26} : x$.

Given the outputs of the four functions at each edge on the change graph, α computes the overall costs of each path from $\mathcal{PAT}_\alpha$ to the set of argumentation theories

$\mathcal{AS}$	$\pm\phi$	Δ_A	Δ_S	Γ_A	Γ_S
$\mathcal{AS}_1$	$+y$	$\emptyset$	$\{\mathcal{A}_{14}, \mathcal{A}'_1\}$	$\{\mathcal{A}_{25}\}$	$\emptyset$
$\mathcal{AS}_1 + y$	$-z$	$\{\mathcal{A}_2\}$	$\emptyset$	$\emptyset$	$\emptyset$
$\mathcal{AS}_1$	$-a_2$	$\{\mathcal{A}_4, \mathcal{A}_{14}, \mathcal{A}_{21}\}$	$\emptyset$	$\emptyset$	$\{\mathcal{A}_8, \mathcal{A}_{17}\}$
$\mathcal{AS}_1$	$+x$	$\emptyset$	$\{\mathcal{A}_3, \mathcal{A}_{14}\}$	$\{\mathcal{A}_{26}\}$	$\emptyset$
$\mathcal{AS}_1 + x$	$-w_1$	$\{\mathcal{A}_1, \mathcal{A}_{13}\}$	$\emptyset$	$\emptyset$	$\emptyset$

Table 6.3: Outputs of the functions for measuring minimal change

$\mathcal{PAT}_\alpha \dot{-} \{\mathcal{A}_8\}$:

$$V(\mathcal{PAT}_\alpha, (\mathcal{PAT}_\alpha + (y, \mathcal{AS}_1)) - (z, \mathcal{AS}_1 + y)) =\mid \{\mathcal{A}_2\} \cup \{\mathcal{A}_{14}, \mathcal{A}'_1\} \cup \{\mathcal{A}_{25}\} \mid= 4$$
$$V(\mathcal{PAT}_\alpha, \mathcal{PAT}_\alpha - (a_1, \mathcal{AS}_1)) =\mid \{\mathcal{A}_4, \mathcal{A}_{14}, \mathcal{A}_{21}\} \cup \{\mathcal{A}_8, \mathcal{A}_{17}\} \mid= 5$$
$$V(\mathcal{PAT}_\alpha, (\mathcal{PAT}_\alpha + (x, \mathcal{AS}_1)) - (w_1, \mathcal{AS}_1 + x)) =$$
$$\mid \{\mathcal{A}_1, \mathcal{A}_{13}\} \cup \{\mathcal{A}_3, \mathcal{A}_{14}\} \cup \{\mathcal{A}_{26}\} \mid= 5$$

id	pl	loc	t	C_{pl}
...	...	...	...	...
13	α	$retract(a, \{b\})$	10	$C_\alpha^{11} \backslash \{a\} \cup \{b\}$
12	α	$retract(\lceil a_1, a_2 \Rightarrow a \rceil, \{y\})$	10	$C_\alpha^{13} \backslash \{\lceil a_1, a_2 \Rightarrow a \rceil\} \cup \{y\}$

Table 6.4: Dialogue progression with α retracting

Thus it can be seen that when α retracts a, the minimal way of doing so is to retract the rule $\lceil a_1, a_2 \Rightarrow a \rceil$, justifying it with y. The subsequent progression of the dialogue is shown in Table 6.4.

This example has illustrated two important principles. The first is that considering argument acceptability is important when determining minimal change in a system of argumentation based on Dung's abstract theory. Had acceptability not been considered, the costs of each change would have been, respectively, 3, 3 and 4, meaning that retracting $\lceil a_1, a_2 \Rightarrow a \rceil$ (along with making the associated justification with y) or a_2 would have been numerically equal. Secondly, the example also shows that minimal change is not always represented by the smallest number of changes. The minimal change identified involves two steps — first the addition of y, then the removal of z so as to maintain well-formedness

and reach the goal with 4 changes overall. Removing only a_2 also reached the goal of the contraction, but did so with a total of 5 changes.

6.3 Lying

Lying can be considered a fundamental human behaviour. Types of lie can range from those which are fairly innocent and trivial (such as telling a child of the existence of Santa Claus), to those with serious consequences (such as a politician lying while in office). Lying is an intrinsic part of specific types of dialogue, such as negotiation, and of specific contextual constraints of dialogue, such as those imposed by social norms and those imposed by the need for privacy.

Sakama et al. [2010] provide a logical account of lying, with three types of dishonesty specified — lying, which is the process of uttering a statement believed to be false; bullshit, which is the process of uttering a statement which is grounded in neither believed truth nor believed falsity; and deceit, which is the process of uttering some information believed to be true, but withholding certain others with the intention that the hearer draws a false conclusion. A main idea in Sakama et al. [2010] is that a liar wishes to keep their dishonesty as small as possible, since by doing so it is easier to maintain. The idea of dishonesty being kept small is strikingly similar to the notion of minimal change in belief revision and, therefore, the theory of argument revision presented in Chapter 5. In very broad terms, a lie is maintained by ensuring that anything that could potentially expose it is not uttered at a future point in the dialogue; this consequence of lying is similar to the consequences of a belief or argument revision process — through uttering the lie, the speaker must in effect, for the purposes of the dialogue at hand, update their beliefs to accommodate it, otherwise they risk being exposed.

The application of argument revision techniques to assist with lying in a dialogue using SPD provides two advantages to a participant:

1. Identification of a “minimal” lie — if more than one potential lie exists to achieve

the same outcome, which one has the least impact on remaining beliefs, with respect to ensuring the lie isn't exposed?

2. Maintaining a lie — if and when a lie is uttered, the revision process used in determining its minimal status will have identified what would be lost and gained in the participant's beliefs and thus what the participant can no longer publicly state, or may have to publicly concede to avoid the lie being exposed

Furthermore, a participant that is open to lying can also assess whether or not lying is minimal with respect to retracting a commitment — that is, does uttering a "minimal" lie have less of an impact than the minimal retraction?

6.3.1 Why lie?

At least two possible scenarios exist where a participant in a dialogue would choose to lie. One is where it possesses no acceptable argument that defeats a previous argument from its opponent; the other is where it does possess acceptable defeaters, but considers them private (see Chapter 4, p.53, Section 4.2.1), thus effectively placing itself in the former scenario, because in the context of a dialogue, a private defeater takes on the same status as an unacceptable one, in that the participant will not communicate it.

An third, and even stronger case for lying can be made when a participant possesses two defeaters of their opponent's argument: one unacceptable and another acceptable, but private. To illustrate this, consider the following knowledge base and rules in an argumentation system, where a government minister is attempting to convince a political opponent that they should not be spending money on welfare. The real reason is that the government thinks a war is about to happen that will need to be paid for — but they do not want to share this information. Instead, the position to not spend money on welfare is justified by saying that money needs to be spent on education, despite the education department being under-budget.

- $\mathcal{K}_p = \left\{ \; have\_money, education\_under\_budget, war \; \right\}$

- $\mathcal{R}_d = \left\{ \begin{array}{l} r1 : war, have\_money \Rightarrow pay\_for\_war \\ r2 : have\_money \Rightarrow pay\_for\_education \\ r3 : have\_money \Rightarrow pay\_for\_welfare \\ r4 : education\_under\_budget \Rightarrow \neg pay\_for\_education \end{array} \right\}$
- $P_{\mathcal{L}} = \{pay\_for\_war\}$

And with:

- $r2 < r1, r3 < r1, r2 < r4, r3 < r2$
- $\overline{pay\_for\_war} = \{pay\_for\_education, pay\_for\_welfare\}$
- $\overline{pay\_for\_education} = \{pay\_for\_war, pay\_for\_welfare\}$
- $\overline{pay\_for\_welfare} = \{pay\_for\_war, pay\_for\_education\}$

The arguments in the system are:

$\mathcal{A}_1$: $have\_money$ $\mathcal{A}_2$: $education\_under\_budget$

$\mathcal{A}_3$: war $\mathcal{A}_4$: $\mathcal{A}_3, \mathcal{A}_1 \Rightarrow_{r1} pay\_for\_war$

$\mathcal{A}_5$: $\mathcal{A}_1 \Rightarrow_{r2} pay\_for\_education$ $\mathcal{A}_6$: $\mathcal{A}_1 \Rightarrow_{r3} pay\_for\_welfare$

$\mathcal{A}_7$: $\mathcal{A}_2 \Rightarrow_{r4} \neg pay\_for\_education$

Using the last-link principle, argument preferences are $\mathcal{A}_5 \prec \mathcal{A}_4$; $\mathcal{A}_6 \prec \mathcal{A}_4$; $\mathcal{A}_6 \prec \mathcal{A}_5$; $\mathcal{A}_5 \prec \mathcal{A}_7$. The set of private arguments is $P_{Args} = \{\mathcal{A}_3, \mathcal{A}_4\}$.

The resultant abstract framework, evaluated under grounded semantics, is shown in Figure 6.5, with acceptable arguments marked with a tick (✓) and private arguments surrounded by a dashed box. For clarity, islands (those arguments with no attack interactions) are omitted.

The politician's beliefs are, therefore,

$B = \{have\_money, war, education\_under\_budget, pay\_for\_war\}$.

Since $\mathcal{A}_3$ is a private argument, the only remaining defeater of $\mathcal{A}_6$ is $\mathcal{A}_5$. However, regardless if $\mathcal{A}_3$ is considered, $\mathcal{A}_5$ is "out" because it is defeated by $\mathcal{A}_7$; thus, no honest defeater of $\mathcal{A}_6$ can be offered, without sharing a private argument.

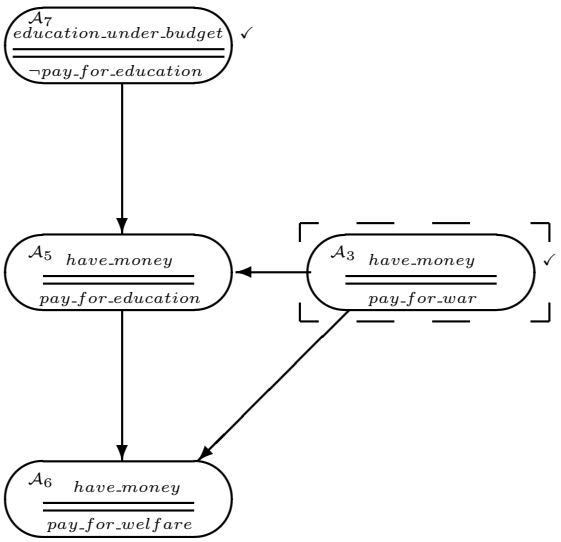

Figure 6.5: Abstract framework for the politician example

What this example demonstrates is a situation in which a participant has the relevant beliefs to defeat their opponent's argument, but cannot share them for privacy reason. They do, however, have an awareness of another argument whose conclusion they do not believe, but claiming would result in the same outcome. If a participant's private arguments lead it to lie, two questions arise: firstly, what, if any, lie should it choose; and secondly how can it be maintained?

In the remainder of this section, we first provide a specification of dishonesty in terms of the ASPIC$^+$ framework and Argument Revision, before using the running example to illustrate the concepts presented.

6.3.2 Characterisation of lying in dialogue

Before answering the questions of what lie to choose and how to maintain it, we first need to characterise what it means to lie, in terms of SPD. Sakama et al.'s [2010] account of lying describes three forms of dishonesty — to lie is to communicate a believed-false statement; to bullshit (henceforth referred to as bluffing) is to communicate a statement whose truth is unknown and to deceive is to communicate (true) information with the intention that the hearer draw an inaccurate conclusion. Since the purpose of this chapter

is to show how argument revision can be used as a tool for supporting dishonesty, attention will be given only to lying and bluffing.

Definition 6.3.1 *(Lie)*

A locution $claim_\alpha(\{\varphi_1, \ldots, \varphi_n\})$ *is a* ***lie*** *if for* $\varphi = \{\varphi_1, \ldots, \varphi_n\}$*:*

- $\varphi \subseteq \bigcup_{\mathcal{A} \in \mathcal{A}rgs(\mathcal{PAT}_\alpha)} Conc(\mathcal{A})$
- $Cl_{R_s}(B_\alpha) \cap \varphi = \emptyset$

That is, α lies if it claims formulae for which there exist an arguments in $\mathcal{PAT}_\alpha$, but it does not believe the formulae, because it possesses no acceptable arguments for them.

Definition 6.3.2 *(Bluff)*

A locution $claim_\alpha(\{\varphi_1, \ldots, \varphi_n\})$ *is a* ***bluff*** *if* $\bigcup_{\mathcal{A} \in \mathcal{A}rgs(\mathcal{PAT}_\alpha)} Conc(\mathcal{A}) \cap \{\varphi_1, \ldots, \varphi_n\} = \emptyset$

That is, α bluffs if it claims formulae for which there exist no arguments in $\mathcal{PAT}_\alpha$ (and also, by extension, it does not believe).

Recall that SPD allows a participant to justify a retraction by possibly providing a contrary or contradictory formula to a formula they have already claimed. This does not preclude dishonesty, however, with the distinction between justified retraction and dishonesty being found in the implicit intent in the locutions.

In a retraction, the participant is offering an explanation as to why it might be the case that what they had previously stated, and are now retracting, is no longer the case. With dishonesty, the participant makes an ordinary claim like any other, with the intention that their opponent accept the dishonest content.

6.3.3 Argument revision and dishonesty

When a participant in a dialogue is dishonest, they claim, and thus become committed to, something they do not believe. However, to avoid the dishonesty being exposed, the

participant must ensure that what they subsequently communicate is consistent with respect to the dishonest claim.

The account of lying provided by Sakama et al. [2010] stipulates that an agent should not tell an "unnecessarily strong" lie, with strength being measured in terms of the impact on other beliefs. Adhering to this principle makes the lie easier to maintain, because with fewer beliefs affected, the agent is less likely to stumble. This idea of a lie having a minimal impact is strikingly similar to the belief and argument revision concepts of minimal change.

Further connections and parallels exist between dishonesty and belief revision, to the extent that lying can be considered a form of belief revision, and thus if beliefs are based on a system of argumentation, argument revision. If Alice lies to Bob, Alice communicates information that they themselves do not believe, but intends that Bob believes they do believe it. It is in this intention that the connection to belief revision is found — while Alice has not revised her actual beliefs (i.e. she does not believe the statement), she has revised the external *perception* of her beliefs. In order to maintain the lie, Alice must continue to present a belief state that is both consistent and logical with respect to the lie itself.

The model of argument revision presented in Chapter 5 can assist a dialogue participant in not only selecting a minimal lie (through use of the minimal change calculation), but also in maintaining it. As with retraction of commitments, a participant will not explicitly revise their underlying personal argumentation theory when lying, because by definition they do not believe the lies they claim. However, a difference between retraction and dishonesty is that the latter needs maintained to avoid exposure and so the participant needs to remain aware of the epistemic state they present. The outputs of the functions for measuring minimal change assist in this regard.

The outputs of the argument drop and acceptability drop functions identify arguments that, if the lie is accommodated, are lost or become unacceptable, but in either case cannot be considered by the participant when establishing their communicable beliefs. A

participant already has the ability to disregard acceptable arguments, and hence beliefs, through the use of private arguments; thus, adding the outputs of the argument and acceptability drop functions to the set of private arguments will allow those arguments to be ignored when choosing a dialogue move.

Formally, given a formula removal or addition performed as part of an argument revision process to support lying, we stipulate that $\Delta^{\pm}_{A}(\phi, \mathcal{AS}) \cup \Delta^{\pm}_{S}(\phi, \mathcal{AS}) \subseteq P_{\mathcal{A}rgs}$.

Arguments that are gained, or gain acceptability have the opposite effect — the participant, depending on the course of the dialogue, may need to appear to believe the conclusions of those arguments. While there is no existing means of maintaining awareness of these arguments (unlike the set of private arguments), the principle is the same; a participant needs a set for keeping track of arguments that are either gained or gain acceptability when a lie is accommodated in their personal argumentation theory.

In terms of the process of Argument Revision, its execution is no different to any other, whether it be for retraction or some other application. However, what does need to be identified is what the process is performed with respect to. For both types of dishonesty (lying and bluffing), the goal is for the participant to find acceptable an argument whose conclusion is the formula upon which the lie was based.

Recall from Chapter 5 the specifications of the two forms of argument revision (contraction, Section 5.2.2, p.76; and expansion, Section 5.2.3, p.76). The goal of contraction is to remove or make unacceptable a given set of arguments, while the goal of expansion is to add or make acceptable a given set of arguments. Dishonesty, therefore, represents an **expansion**, with the participant looking to add and/or make acceptable arguments that conclude the lie.

Figure 6.3.3 shows the final stages of dialogue move selection, incorporating lying.

6.3.4 Running example

In the previous section, an example was provided to illustrate why a dialogue participant would choose to be dishonest. In this section, we return to the running example presented

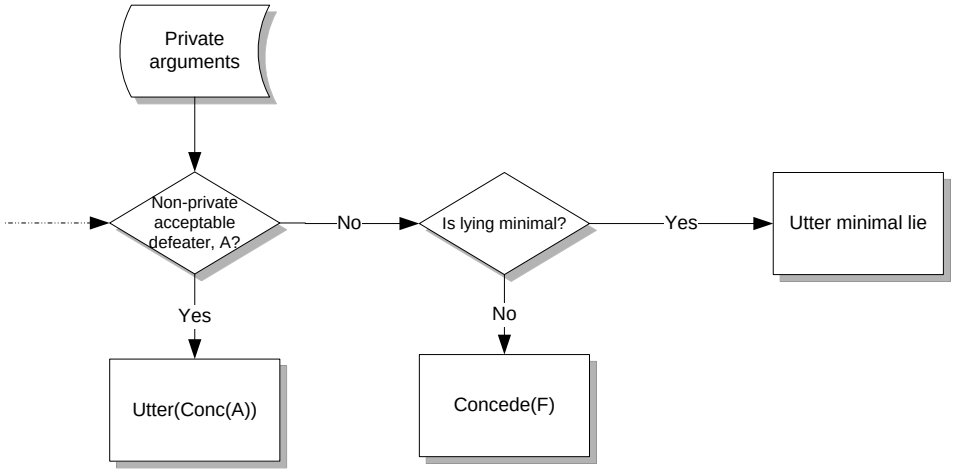

Figure 6.6: Dialogue move selection incorporating lying (initial steps omitted)

in Chapter 4, and continued in section 6.2.2 to show how dishonesty can be integrated into the overall process of choosing a dialogue move.

In section 6.2.2, α identified that conceding β's claims, then retracting $\lceil a_1, a_2 \Rightarrow a \rceil$ and justifying it with y represented the minimal change to their existing commitments and remaining beliefs. However, now that dishonesty is being considered as a means of defending a position, both α's original concession and then all subsequent moves need to be re-evaluated with respect to the possible lies or bluffs that could help them defend their position.

In Figure 6.7, a reminder of the abstract framework from $\mathcal{PAT}_\alpha$ is provided, which now also incorporates β's argument for d (where, having evaluated the framework under grounded semantics, acceptable arguments are marked with a tick; private arguments are surrounded with a dashed box). Notice that α has an acceptable defeater of β's argument for d, but it is private; they possess one other defeater for d, the argument for g, but consider it it unacceptable. To assess the effects of making the argument for g acceptable, α can perform an argument expansion with respect to g.

When an argument already exists in a system, but is unacceptable, the goal of argument expansion is to make it acceptable. To make the argument for g acceptable, α can either introduce a defeater of, or completely remove the argument for h, the sole defeater of g. Table 6.5 shows the possible ways in which this can be achieved.

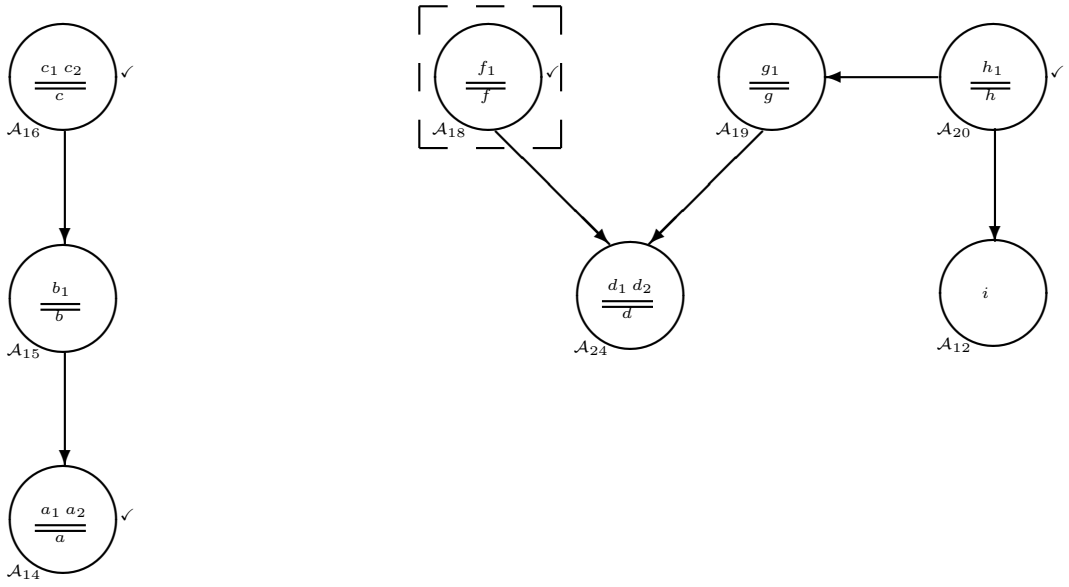

Figure 6.7: Framework from $\mathcal{PAT}_\alpha$, incorporating β's argument for d

$\mathcal{AS}$	$\pm\phi$	Δ_A	Δ_S	Γ_A	Γ_S
$\mathcal{AS}_1$	$-h_1$	$\{\mathcal{A}_{11}, \mathcal{A}_{20}\}$	$\emptyset$	$\emptyset$	$\{\mathcal{A}_{12}, \mathcal{A}_{19}\}$
$\mathcal{AS}_2$	$-\lceil h_1 \Rightarrow h \rceil$	$\{\mathcal{A}_{20}, \mathcal{A}'_7\}$	$\emptyset$	$\emptyset$	$\{\mathcal{A}_{12}, \mathcal{A}_{19}\}$
$\mathcal{AS}_2$	$-\lceil h \in \overline{g} \rceil$	$\{\mathcal{A}'_{13}\}$	$\emptyset$	$\emptyset$	$\{\mathcal{A}_{19}\}$

Table 6.5: Outputs of the functions for measuring minimal change in $\mathcal{PAT}_\alpha \dot{+} \{\mathcal{A}_{19}\}$

All three possible methods of performing the expansion rely on only one step, thus the change graph as a structure is relatively trivial, with each expansion represented by a single edge. This can be seen in Figure 6.8. However, the measures for minimal change still allow the least costly lie to be identified, and whether or not uttering it is preferable (with respect to minimal change) to conceding then retracting.

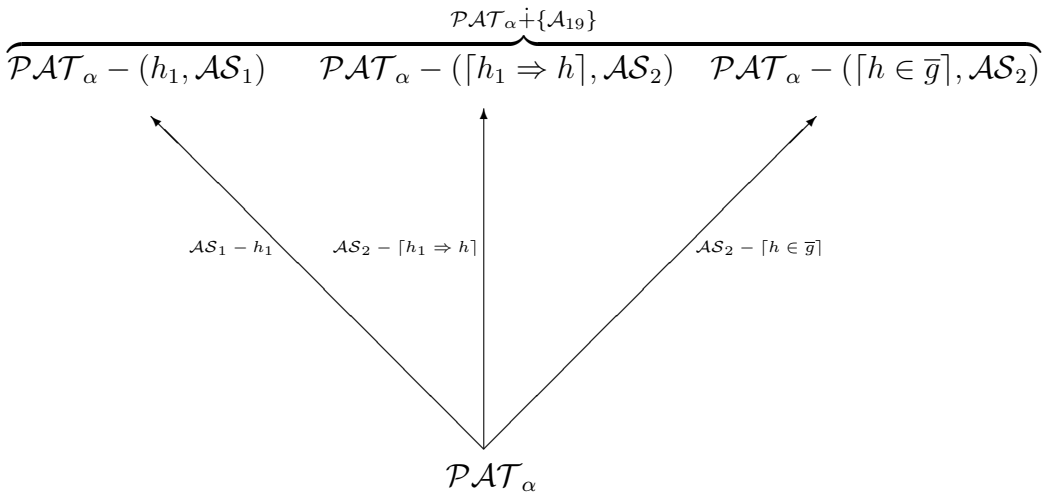

Figure 6.8: Change graph for $\mathcal{PAT}_\alpha \dot{+} \{\mathcal{A}_{19}\}$

The cost of each path is:

$$V(\mathcal{PAT}_\alpha, \mathcal{PAT}_\alpha - (h_1, \mathcal{AS}_1)) =\mid \{\mathcal{A}_{11}, \mathcal{A}_{20}\} \cup \{\mathcal{A}_{12}, \mathcal{A}_{19}\} \mid = 4$$
$$V(\mathcal{PAT}_\alpha, \mathcal{PAT}_\alpha - (\lceil h_1 \Rightarrow h \rceil, \mathcal{AS}_2)) =\mid \{\mathcal{A}_{20}, \mathcal{A}'_7\} \cup \{\mathcal{A}_{12}, \mathcal{A}_{19}\} \mid = 4$$
$$V(\mathcal{PAT}_\alpha, \mathcal{PAT}_\alpha - (\lceil h \in \overline{g} \rceil, \mathcal{AS}_2)) =\mid \{\mathcal{A}'_{13}\} \cup \{\mathcal{A}_{19}\} \mid = 2$$

If α chooses to lie by considering the argument for g acceptable, the minimal way of doing this is to ignore the contrary (attack) between h and g. The continuation of the dialogue in this situation is in Table 6.6.

id	pl	loc	t	C_{pl}
...	...	...	...	...
13	α	$claim(g)$	10	$C_\alpha^{11} \cup \{g\}$
14	β	$why(g)$	11	C_β^{12}
15	α	$claim(\{g_1, \lceil g_1 \Rightarrow g \rceil\})$	14	$C_\alpha^{13} \cup \{g_1, \lceil g_1 \Rightarrow g \rceil\}$

Table 6.6: Continuation of dialogue when α lies

Retracting vs. lying

Recall from section 6.2.2 that α identified that justifying their retraction of a by claiming y (an undercutter to the argument for a), then removing z was minimal, resulting in four changes overall to what they can communicate. However, when examining the effects of lying, claiming g and making private belief in $h \in \overline{g}$ resulted in two changes overall. Therefore, lying is optimal, with respect to minimality.

The example has therefore demonstrated that argument revision can assist with not only determining the minimal retraction or lie, but also in determining which dialogue move (i.e. $retract$ or dishonest $claim$) to select in the first place.

6.4 Summary

In this chapter, the model for argument revision presented in Chapter 5 has been applied to the dialogue framework specified in Chapter 4 to show how it can assist a participant in selecting both their next move and determine its content.

We first continued the running example from Chapter 4 by showing α as conceding β's argument for d, which caused their commitment store to become inconsistent and their original argument (a) to be undefended in the shared argumentation theory. This led β to force resolutions of the inconsistency and lack of defence. The inconsistency was resolved through retracting a formula at the source of it, while the lack of defence of a was resolved through retracting it.

However, α also had to retract further arguments to ensure that a was not in the *closure* of their commitment store. Three possible means of doing so were identified, with

argument revision techniques being employed to choose between them. In doing so, two important principles were made explicit:

1. Argument acceptability is an important consideration when determining minimal change in a system of argumentation based on Dung's abstract theory.

2. The smallest number of explicit changes to an argumentation theory does not necessarily represent the minimal change overall — in the example, one change resulted in five changes overall; two changes resulted in four.

In addition to applying argument revision to retraction, this chapter also presented an alternative to conceding in the first place — dishonesty. A participant could choose to be dishonest instead of conceding a position to their opponent and, ultimately, being forced to resolve inconsistencies in their commitment store and a lack of defence of the position. A participant may choose to be dishonest for two reasons; either they wish to defend their position at all costs, or they do have an argument that can act as a defence, but (using some unspecified criteria) they have decided to keep it private. They therefore look to an unacceptable argument to provide a defence.

A formal specification of two types of dishonesty — lying and bluffing — was provided, influenced by the account specified by Sakama et al. [2010]. A lie is the utterance of a statement where the participant believes a contrary statement, while a bluff is the utterance of a statement that is grounded neither in truth nor falsity.

Argument revision was then shown to be a means through which dishonesty can be supported, from the initial decision to be dishonest, to maintaining it to avoid detection. The initial decision is made through using the principle of minimal change to compare dishonesty to the effects of conceding, while maintenance is achieved through using the outputs of the four functions for measuring minimal change as a way of the participant being aware of what they cannot publicly become committed to, or may have to appear to be committed to, as a result of the lie or bluff.

Chapter 7

Conclusions

7.1 Summary

This thesis has presented a model for argument revision in the ASPIC$^+$ system of structured argumentation. The model is influenced by the AGM theory of belief revision [Alchourrón et al., 1985] in that it is driven by a concept of minimal change. However, a key difference between the AGM theory and the present model is the discarding of an entrenchment ordering over beliefs. Instead, minimal change is arrived at by considering measurable effects on the system when a revision process is performed.

Two types of argument revision were specified: argument contraction, where the goal is to remove or make unacceptable a set of arguments in an argumentation theory; and argument expansion, where the goal is to add or make acceptable a set of arguments in an argumentation theory. Both processes are similar to their belief revision counterparts, except that the goal of each does not rely on a consistent argumentation theory (since consistency is handled by argument evaluation), but instead requires that the resultant argumentation theory be well-formed.

It was identified that, at the basic level, an argumentation theory can be revised by modifying the knowledge base in an argumentation system to add or remove premises. Two functions, formula addition and formula removal were defined to perform these actions.

Formula addition and removal have no regard for the overall goal of the revision, but instead take as input an argumentation system and set of formulae, and output a new argumentation system with the formulae added to, or removed from, the knowledge base. To achieve the goal of the revision, if it does not happen through a single action, further additions and removals can take place.

When performing an addition or removal, the effects on the system are captured through four functions: argument drop, argument gain, acceptability drop and acceptability gain. These functions identify the arguments affected when a formula is added to or removed from the knowledge base in an argumentation system. Where multiple changes are required to reach the goal of the revision, the measures are kept separate until the goal is achieved, because it is possible that the effects of one change are subsequently undone by another — for instance, an argument that loses acceptability may be reinstated in a later change in the same revision process. A structure called a change graph was defined to model the possible changes, with an edge cost function being used to apply the measures to each edge.

As well as being able to revise arguments on their premises, it was also identified as being desirable to revise on rules, preferences and contraries. Revising on rules provides an additional means by which arguments can be completely removed from a system, while revising preferences and contraries allows attacks and defeat to be modified so as to make arguments (un)acceptable. To allow such changes, meta-level extensions to the ASPIC$^+$ framework were defined, where every rule, preference and contrary in an ASPIC$^+$ argumentation system are represented by (possibly atomic) arguments at the meta-level. Revising these meta-level arguments (through formula additions and removals at the meta-level, using the same functions as already defined) is reflected in the object-level components and, by extension, arguments.

The specification of meta-argumentation in ASPIC$^+$ differs from previous specifications, such as that by Weide and Dignum [2011], in that it requires meta-level arguments for all non-knowledge base components at the object-level. This results in all aspects of an argumentation theory having argumentative backing. It also allows for a better

characterisation of strict and defeasible rules, such that an object-level rule r is strict iff at the meta-level, no formula is declared a contrary of r; otherwise, it is defeasible. This, in turn, allows for a more principled definition of undercutting that does not rely on the rules, or their labels, needing to be represented in the object-language: a defeasible rule is undercut if there is an acceptable argument for any formula which is declared a contrary of it.

To demonstrate the application of argument revision in dialogue, a protocol for a simple persuasion dialogue, SPD was specified. Using this protocol, a running example showed how a participant in the dialogue can use argument revision when faced with an inability to respond to an opponent's claim. The participant could either choose to concede the claim, and be forced to give up previous commitments, or they could choose to be dishonest.

Argument Revision allows a participant to decide between conceding (and possibly retracting) and being dishonest. Using the measures of minimal change, the participant can assess whether it is minimal to concede, then retract, or to continue to defend their position by being dishonest. If they choose to concede and retract, Argument Revision then allows the minimal set of retractions to be identified that allow the commitment store to remain indirectly consistent. If they choose to be dishonest, Argument Revision allows for both the selection and maintenance of a lie. A minimal lie is first chosen, using the determination of minimal change, with the outputs of the four functions being used to identify arguments that end up "off-limits" at a future point in the dialogue if the participant does not wish for the lie to be exposed.

7.2 Contributions

The contributions of this thesis are threefold: (i) meta-level extensions to the ASPIC$^+$ framework; (ii) a model of argument revision that does not rely on a pre-determined entrenchment ordering; and (iii) the decoupling of argument dynamics from the framework to which they apply.

7.2.1 Meta-level extensions to the ASPIC$^+$ framework

Chapter 3 presents a specification of meta-level extensions to the ASPIC$^+$ framework. This specification provides that given an object-level argumentation system, every non-knowledge base component of that system (rule, preference and contrariness) is represented in the meta-language, with the associated meta-argumentation system containing an argument that concludes that meta-level representation.

Requiring rules, preferences and contrariness to be represented as (meta-level) arguments provides advantages for both argument revision and the use of ASPIC$^+$ in general. In an argument revision context, because a meta-argumentation system is structurally no different to one at the object-level (with the main difference being the logical language used to instantiate the system), a single method of argument revision based on knowledge base modification can be used to revise arguments based on premises, rules, preferences and/or attacks.

In a broader context, the specification allows for a refined definition of undercutting that does not rely on rules being named in the object language. Informally, an inference rule is strict iff it holds without exception; otherwise, it is defeasible. An exception is some situation in which a certain rule cannot be applied: "in situation X it is the case that rule r does not hold". It is not required for X to be true — all that is required is the relation between X and the rule it undercuts.

Using this characterisation, we can view X as a contrary of the rule, which can be represented using our specification of meta-argumentation. A rule r is strict iff there does not exist an argument $\lceil X \in \overline{\lceil r \rceil} \rceil$ at the meta-meta-level; otherwise, it is defeasible.

The specification of meta-level extensions to ASPIC$^+$ ensures that every component (rule, preference, contrary) of an argumentation theory is represented by a (meta-) argument, which provides advantages not only in Argument Revision, but in other applications as well; with every component represented as an argument, they can be reasoned about in the same way as ordinary, object-level arguments. For instance, an argument for a rule can be presented, providing justification for why this rule holds.

7.2.2 Revision without entrenchment

When using the AGM theory to revise beliefs, a determination of minimal change is based on a qualitative entrenchment ordering over a belief set. Such an ordering limits the ability of an agent in a dynamic environment to evaluate information it has accepted post-deployment (and hence post-determination of the entrenchment ordering).

Nevertheless, the idea of minimal change itself is reasonable — update beliefs with the minimal impact on those that remain. Thus, the model for argument revision presented in Chapter 5 incorporates minimal change in assessing argument dynamics, but relies on a new, justifiable means of determining it.

Instead of requiring a pre-determined entrenchment ordering, minimality is evaluated based on measurable effects on the system. Using the ASPIC$^+$ framework, four measures were identified for a change operation: argument loss, those arguments completely removed from the system; acceptability loss, those arguments that remain in the system but lose acceptability; argument gain, those arguments that are newly added to the system; and acceptability gain, those arguments that were previously in the system and have gained acceptability.

A *change graph* was defined as a means of identifying all possible methods of performing a certain revision, with a cost function being used to apply the measures of minimal change to each edge. An operator to compute a path cost, which considers the net drops and gains for each measure, was also defined so as to determine what constitutes the overall minimal change.

Removing the need for an entrenchment ordering over beliefs (or arguments), and instead determining minimal change based on quantifiable, measurable effects a change has on the system, allows an agent in a dynamic environment to evaluate the importance of new information that it was not aware of when its entrenchment ordering was specified.

7.2.3 Decoupling dynamics from the system

The model for argument revision presented in this thesis does not rely on a new argumentation framework, or one tailored specifically to include dynamics, but instead uses an existing system and remains completely external to it. The model is defined by a set of processes that take as input an argumentation theory and the required changes, and output a set of new, possible argumentation theories where the changes have taken place.

This decoupling provides both theoretical and application-based advantages. From a theoretical standpoint, frameworks for argumentation, such as the ASPIC$^+$ framework, have several established properties. Redefining ASPIC$^+$ to have an intrinsic model of dynamics would require either to prove the properties still hold, or show that they don't. Furthermore, extensions to the framework can, in principle, make use of the revision model without needing to be redefined on the basis of a new version of the framework that incorporates dynamics.

In terms of applications, decoupling the model of dynamics allows it to be used as a tool for evaluating a framework, as opposed to it being an actual part of the framework. In a dialogical context, this allows a dialogue to be based on the framework, with the revision model used as a strategic tool by some, but not necessarily all, of the participants. In dialogues based on frameworks where dynamics intrinsic to the framework, every participant has the ability to assess dynamics, based on the same criteria, removing the strategic edge.

7.3 Future work

Directions for future work fall into two categories; theoretical extensions, and the exploration of further applications, from both dialogical and wider standpoints.

7.3.1 Theoretical work

Several areas offer the potential for further work on the theoretical models presented.

Meta-argumentation

The specification of meta-argumentation outlined in Chapter 3 provides only a set of bijections to define a meta-argumentation system. The process of elevating object-level components to meta-level arguments, and vice versa, is left implicit for the time being, however for the model to be of further use, it will need to be made explicit.

Moving from the object-level to the meta-level is relatively trivial, because it involves a flattening of the components into formulae of the meta-language, and subsequently their type (rule, preference, contrary) is unimportant. However, if we start with meta-arguments for the components, translating them into their relevant object-level type is more complex, because unless assumptions are placed on the meta-language regarding the internal structure of formulae, there is no formal method of distinguishing a rule from a preference or a contrary.

Argument revision

The model for argument revision has scope for further refinement through considering further features of the ASPIC$^+$ framework. While preferences between arguments can be revised (on their meta-level representations), they are not presently considered when choosing between possible revision methods. Incorporating preferences into the model could, for instance, assist in situations where two possible methods yield the same number of changes; if one particular method perhaps drops an argument that is more preferred to a dropped argument in another method, the former could be chosen.

A different form of preference would be to place a preference ordering over the functions for argument revision (argument drop and gain, and acceptability drop and gain). Such an ordering could see, for example, argument drops being given a higher priority than argument gains, and so a method with more gains than drops (but the same number of overall changes) is chosen.

A further extension to the model is represented by a more general possible extension to the ASPIC$^+$ framework, which is the consideration of the strength of defeasible rules,

based on exceptions. Presently, the only formal role of exceptions is to determine whether or not an inference rule should be applied (if there is an acceptable argument for an exception it should not; otherwise, it should). However, it is possible to use exceptions in a determination of rule strength, which in turn can have applications not only in argument revision, but also in developing an automatic method of arriving at rule preferences.

Broadly speaking, a rule's strength is inversely-proportional to the number of exceptions to it (regardless if the criteria are met for the exception resulting in the rule not being applicable) — so a rule with n exceptions is stronger than a rule with $> n$ exceptions. In argument revision, this would be useful when revising an argument with multiple rules; although the exact application requires further investigation, we envisage that given a numerical value of strength (based on the number of exceptions) that value could be used as a scaling factor on the number of changes when removing or introducing an exception to that rule.

7.3.2 Applications

Argument web

An obvious direction for future work in terms of applications is to exploit the connections between ASPIC$^+$ and the Argument Interchange Format [Bex et al., 2010, 2013] and apply the argument revision model to real argument data. Exploring the dynamics of real data would represent a valuable addition to tools, such as Arvina [Lawrence et al., 2012a], which employ dialogue-based exploration of argument space [Reed and Wells, 2007] using mixed-initiative argumentation. When engaged in a debate using an Arvina-style tool, a user would be able to employ argument revision as a strategic tool in the same way as autonomous agents.

Dialogue

While this thesis has used argument revision to assist in a persuasion dialogue, it also has potential applications in other dialogue types. For instance, in a negotiation dialogue, the aim of each participant is to get the best outcome for themselves. The application of argument revision could allow a participant to assess what utterances, whether honest or dishonest, allow this to be achieved.

There is also a potential application in co-operative dialogue domains. In their work on teamwork in multi-agent systems, Dunin-Kęplicz and Verbrugge [2010] use dialogue as a means through which agents can form teams. They specify a new dialogue type, *persuasion with respect to motivational attitudes*, which arises from a situation of one party wishing to achieve φ, while other parties have a conflicting intention (e.g. ψ, which is inconsistent with φ), or simply have no positive motivation towards achieving φ. The main goal of this dialogue is the resolution of conflict to result in a stable collective intention. Argument revision can play a part in helping resolve the conflict through its use as a tool by the other parties in assessing the effect of taking on the intention for φ.

Similar situations exist in intrinsically co-operative dialogues, such as inquiry [Black and Hunter, 2007, 2009], where two or more participants work together to establish a truth. In doing so, one or both may uncover information that is inconsistent with their present beliefs, and so need to evaluate the cost of accommodating it (and sacrificing existing beliefs).

7.4 Validating the research hypothesis

In Chapter 1, the research hypothesis motivating this thesis stated that:

> *A model of argument revision can be developed that is applied to, but remains independent of, an existing system of argumentation, which in turn can be deployed as a strategic tool in multi-agent dialogue.*

This thesis has specified a model for argument revision based on the ASPIC$^+$ system. The model does not redefine the system in any way, but instead uses existing features and properties of the system in defining and describing change operations applicable to the system. To provide a full account of argument dynamics, that considers all elements on an argumentation system (i.e. rules, preferences and contraries as well as elements of the knowledge base), meta-level extensions to the system were defined. However, we argue that these extensions are not specific to argument revision, but are more general and have a use beyond the dynamics of argumentation.

The thesis has also shown how the argument revision model can be used by a participant in a dialogue to determine, in certain challenging situations, their next move and the content of that move. When faced with a scenario in which their position cannot be defended, a participant has the choice of either conceding the opponent's position, then retracting (some of) their own, or being dishonest.

Argument revision assists by first allowing the participant to determine whether being honest (i.e. conceding) is minimal compared to being dishonest. If it is, the same technique is then used to determine the minimal set of retractions required to remain consistent. If being dishonest is identified as minimal, argument revision allows both the minimal dishonest act to be identified, and the dishonesty to be maintained. The latter is achieving by using the measures of minimal change as a means of retaining awareness of the arguments affected by outwardly appearing to believe the dishonest statement.

7.5 Conclusions

This thesis has presented a model for argument revision, inspired by the AGM theory of belief revision and designed to explore the dynamics of an existing system of structured argumentation, and demonstrated how the model can be used as a strategic tool for a participant engaged in a dialogue. While using AGM-style operators and concepts to model the dynamics of an argumentation system of framework is not in itself new, we believe that

modelling the dynamics of an *existing* (and not newly-defined) system is. Furthermore, the model presented here provides an intuitive, justifiable algorithmic determination of minimal change, that does not rely on a pre-determined entrenchment ordering but instead uses measurable effects on the system when a change is performed.

A link between belief revision and the dynamics of dialogue was also identified and explored as an application of the argument revision model. A simple dialogue protocol was specified, based on the ASPIC$^+$ framework for argumentation, and used to illustrate two scenarios where applying argument revision techniques could prove beneficial to a participant engaged in the dialogue. Both scenarios arise from a position of being unable to defend a position against an opponent's argument(s).

The first scenario sees the participant use argument revision and its determination of minimal change as a means of determining the effect of conceding an argument to their opponent. Upon conceding, the participant may subsequently be forced into retracting previous commitments inconsistent with the concession, as well as any other commitments that indirectly (through inference) result in an inconsistency. It was shown that argument revision assists with this process by allowing a participant to evaluate, with respect to minimal change, the effect of different retractions on both existing commitments, and potential future commitments.

The second scenario is aimed at providing an alternative to conceding an argument to an opponent — dishonesty. A dishonest act sees the participant utter in the dialogue something they do not believe, with the intention that their opponent does believe it. A characterisation of dishonesty, based on the formal specification of Sakama et al. [2010], was provided, then uses to show how dishonesty can be characterised as a belief revision, and by extension argument revision, process. When being dishonest, a participant still does not believe the statement or argument at the source of the dishonesty, but must externally appear to do so. The application of argument revision techniques allows for the minimal dishonesty to be determined — that is, a dishonest act that has the least impact on other beliefs. Keeping dishonesty as small as possible reduces the risk of it being exposed and

prevents other beliefs from being “off-limits” later in the dialogue.

These two applications of argument revision in dialogue where then combined to show that applying the techniques can allow a participant to determine whether conceding then retracting or being dishonest has the minimal impact on their beliefs, before choosing exactly which retraction to perform or dishonest act to utter.

Bibliography

C.E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2):510–530, 1985.

L. Amgoud. Five weaknesses of ASPIC+. In *Proceedings of the 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, LNAI, pages 122–131, Verlag, 2012. Springer.

L. Amgoud and C. Cayrol. Integrating preference orderings into argument-based reasoning. In *Proceedings of the First International Joint Conference on Qualitative and Quantitative Practical Reasoning*, pages 159–170, London, UK, 1997. Springer.

L. Amgoud, L. Bodenstaff, M. Caminada, P. McBurney, S. Parsons, H. Prakken, J. van Veenen, and G.A.W. Vreeswijk. Final review and report on formal argumentation system. Deliverable D2.6, ASPIC IST-FP6-002307, 2006.

L. Amgoud, C. Cayrol, and M.C. Lagasquie-Schiex. On the bipolarity in argumentation frameworks. *International Journal of Intelligent Systems*, 23:1062–1093, 2008.

P. Baroni, F. Cerutti, M. Giacomin, and G.R. Simari, editors. *Computational Models of Argument: Proceedings of COMMA 2010*, Desenzano del Garda, Italy, 2010. IOS Press.

T. Bench-Capon and P. Dunne. Value based argumentation frameworks. Technical report, University of Liverpool, 2002.

T.J.M. Bench-Capon and H. Prakken. Using argument schemes for hypothetical reasoning in law. *Artificial Intelligence and Law*, 18:153–174, 2010.

Ph. Besnard and A. Hunter. *Elements of Argumentation*. MIT Press, 2008.

Ph. Besnard, S. Doutre, and A. Hunter, editors. *Computational Models of Argument: Proceedings of COMMA 2008*, Toulouse, France, 2008. IOS Press.

F. Bex. *Arguments, Stories and Criminal Evidence: A Formal Hybrid Theory*. Springer, 2011.

F. Bex, H. Prakken, and C. Reed. A formal analysis of the AIF in terms of the ASPIC framework. In P. Baroni, F. Cerutti, and G.R. Simari, editors, *Proceedings of the Third International Conference on Computational Models of Argument (COMMA 2010)*, pages 99–110, Desenzano del Garda, Italy, 2010. IOS Press.

F. Bex, S. Modgil, H. Prakken, and C. Reed. On logical reifications of the argument interchange format (to appear). *Journal of Logic and Computation*, 2013.

E. Black and A. Hunter. A generative inquiry dialogue system. In E.H. Durfee, M. Yokoo, M.N. Huhns, and O. Shehory, editors, *Sixth International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2007)*, pages 241:1–241:8, 2007.

E. Black and A. Hunter. An inquiry dialogue system. *Autonomous Agents and Multi-Agent Systems*, 19:173–209, 2009.

N. Bulling, C. Chesñevar, and J. Dix. An argumentative approach for modelling coalitions using atl. In I. Rahwan and P. Moraitis, editors, *ArgMaS 2008*, Lecture Notes in Computer Science, pages 197–246. Springer, 2009.

M. Caminada. Semi-stable semantics. In P.E. Dunne and T.J.M. Bench-Capon, editors, *Proceedings of the 1st International Conference on Computational Models of Argument (COMMA 2006)*, pages 121–130, Liverpool, UK, 2006. IOS Press.

M. Caminada. Comparing two unique extension semantics for formal argumentation: Ideal and eager. In *Proceedings of BNAIC 2007*, pages 81–87, 2007a.

M. Caminada. An algorithm for computing semi-stable semantics. In *Proceedings of ECSQARU 2007*, pages 222–234, 2007b.

M. Caminada and L. Amgoud. On the evalutation of argumentation formalisms. *Artificial Intelligence*, 171:286–310, 2007.

C. Cayrol, F.D. de Saint-Cyr, and M. Lagasquie-Schiex. Change in abstract argumentation frameworks: Adding an argument. *Artificial Intelligence Research*, 38:49–84, 2010.

C. Chesñevar, J. McGinnis, S. Modgil, I. Rahwan, C. Reed, G. Simari, M. South, G. Vreeswijk, and S. Willmott. Towards an argument interchange format. *The Knowledge Engineering Review*, 21(4):293–316, 2006.

J. Devereux. *Strategies for Persuasion in Inter-Agent Dialogue*. PhD thesis, University of Dundee, 2011.

J. Devereux and C. Reed. Strategic argumentation in rigorous persuasion dialogue. In *Proceedings of the Sixth International Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2009)*, Lecture Notes in Computer Science, pages 94–113, 2009.

J. Doyle. A truth maintenance system. *Artificial Intelligence*, 12:231–272, 1979.

P. M. Dung. On the acceptability of arguments and its fundemental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

P.M. Dung, P. Mancarella, and F. Toni. Computing ideal sceptical argumentation. *Artificial Intelligence*, 171:642–674, 2007.

B. Dunin-Kęplicz and R. Verbrugge. *Teamwork in Multi-Agent Systems*. Wiley, 2010.

P.E. Dunne and T.J.M. Bench-Capon, editors. *Computational Models of Argument: Proceedings of COMMA 2006*, Liverpool, UK, 2006. IOS Press.

M. A. Falappa, G. Kern-Isberner, and G.R. Simari. Explanations, belief revision and defeasible reasoning. *Artificial Intelligence*, 141:1–28, 2002.

M. A. Falappa, G. Kern-Isberner, and G. R. Simari. Belief revision and argumentation theory. In I. Rahwan and G.R. Simari, editors, *Argumentation in Artificial Intelligence*, pages 314–360. Springer, 2009.

M. A. Falappa, A. J. Garcia, G. Kern-Isberner, and G. R. Simari. On the evolving relation between belief revision and argumentation. *The Knowledge Engineering Review*, 26(1): 35–43, 2011.

M. A. Falappa, G. Kern-Isberner, M. D. L. Reis, and G.R. Simari. Prioritized and non-prioritized multiple change on belief bases. *Journal of Philosophical Logic*, 41(1): 77–113, 2012.

E.L. Fermé and S.O. Hansson. Selective revision. *Studia Logica*, 63(3):331–342, 1999.

A.J. García and G.R. Simari. Defeasible logic programming: An argumentative approach. *Theory and Practice of Logic Programming*, 4(2):95–138, 2004.

P. Gärdenfors. *Knowledge in Flux*. MIT Press, 1988.

P. Gärdenfors. Belief revision: An introduction. In P. Gärdenfors, editor, *Belief Revision*, pages 1–27. Cambridge University Press, 1992.

R.A. Girle. Belief sets and commitment stores. In *Proceedings of the 2nd OSSA conference*, 1997.

R.A. Girle. Positive agnosticism in belief revision. In *Logic and Multi-agent Systems Workshop*, University of Otago, October 2002.

S.O. Hansson. A survey of non-prioritized belief revision. *Erkenntnis*, 50(2-3):413–427, 1999.

P. Krümpelmann, M. Thimm, M.A. Falappa, A. Garca, G. Kern-Isberner, and G.R. Simari. Selective revision by deductive argumentation. In S. Modgil, N. Oren, and F. Toni, editors, *Theory and Applications of Formal Argumentation*, pages 147–162, Barcelona, Spain, 2012. Springer.

J. Lawrence, F. Bex, and C. Reed. Dialogues on the argument web: Mixed initiative argumentation with arvina. In B. Verheij, S. Szeider, and S. Woltran, editors, *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*, pages 513–514, Vienna, Austria, 2012a. IOS Press.

J. Lawrence, F. Bex, C. Reed, and M. Snaith. Aifdb: Infrastructure for the argument web. In B. Verheij, S. Szeider, and S. Woltran, editors, *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*, pages 515–516, Vienna, Austria, 2012b. IOS Press.

I. Levi. Subjunctives, dispositions and chances. *Synthese*, 34:423–455, 1977.

P. McBurney and S. Parsons. Dialogue games in multi-agent systems. *Informal Logic*, 22(3):257–274, 2002.

S. Modgil. Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, 173(9-10):901–1040, 2009a.

S. Modgil. Labellings and games for extended argumentation frameworks. In *Proceedings of the twenty-first International Joint Conference on Artificial Intelligence (IJCAI '09)*, 2009b.

S. Modgil and H. Prakken. Reasoning about preferences in structured extended argumentation frameworks. In P. Baroni, F. Cerutti, and G.R. Simari, editors, *Proceedings of the 3rd International Conference on Computational Models of Argument (COMMA 2010)*, pages 347–358, Desenzano del Garda, Italy, 2010. IOS Press.

S. Modgil and H. Prakken. Revisiting preferences and argumentation. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, pages 1021–1026, 2011.

S. Modgil and H. Prakken. A general account of argumentation with preferences (to appear). *Artificial Intelligence*, 2013.

M.O. Moguillansky, N.D. Rotstein, M.A. Falappa, A.J. Garcia, and G.R. Simari. Argument theory change through defeater activation. In P. Baroni, F. Cerutti, and G.R. Simari, editors, *Proceedings of the 3rd International Conference on Computational Models of Argument (COMMA 2010)*, pages 359–366, Desenzano del Garda, Italy, 2010. IOS Press.

N. Oren. *An Argumentation Framework Supporting Evidential Reasoning with Applications to Contract Monitoring*. PhD thesis, University of Aberdeen, 2007.

S. Parsons and N. Jennings. Negotiation through argumentation — a preliminary report. In *Proceedings of the Second International Conference on Multiagent Systems (ICMAS 1996)*, pages 267–274, 1996.

P. Pilotti, A. Casali, and C. Chesñevar. A belief revision approach for argumentation-based negotiation with cooperative agents (to appear). In *Proceedings of the ninth International Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2012)*, Valencia, Spain, 2012. Springer.

J.L. Pollock. Defeasible reasoning. *Cognitive Science*, 11:481–518, 1987.

H. Prakken. Coherence and flexibility in dialogue games for argumentation. *Journal of Logic and Computation*, 15(6):1009–1040, 2005.

H. Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1(2):93–124, 2010.

H. Prakken and S. Modgil. Clarifying some misconceptions on the aspic+ framework. In B. Verheij, S. Szeider, and S. Woltran, editors, *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*, pages 442–453, Vienna, Austria, 2012. IOS Press.

H. Prakken and G.A.W. Vreeswijk. Logics for defeasible argumentation. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 4, pages 219–318. Kluwer Academic Publishers, 2002.

I. Rahwan and C. Reed. The argument interchange format. In I. Rahwan and G. Simari, editors, *Argumentation in Artificial Intelligence*, pages 383–402. Springer, 2009.

I. Rahwan, F. Zablith, and C. Reed. Laying the foundations for a world wide argument web. *Artificial Intelligence*, 171:897–921, 2007.

C. Reed. Dialogue frames in agent communication. In Y. Demazeau, editor, *Proceedings of the Third International Conference on Multi-Agent Systems (ICMAS 1998)*, pages 246–253, Paris, France, 1998. IEEE Press.

C. Reed and G. Rowe. Araucaria: software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13:961–980, 2004.

C. Reed and C.W. Tindale, editors. *Dialectics, Dialogue and Argumentation: An Examination of Douglas Walton's Theories of Reasoning and Argument*. College Publications, London, 2010.

C. Reed and S. Wells. Dialogical argument as an interface to complex debates. *IEEE Intelligent Systems*, 22(6):60–65, 2007.

C. Reed, S. Wells, J. Devereux, and G. Rowe. AIF+: Dialogue in the argument interchange format. In Ph. Besnard, S. Doutre, and A. Hunter, editors, *Proceedings of the Second International Conference on Computational Models of Argument (COMMA 2008)*, pages 311–323, Toulouse, France, 2008. IOS Press.

R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1-2):81–132, 1980.

N.D. Rotstein, M.O. Moguillansky, M.A. Falappa, A.J. Garcia, and G.R. Simari. Argument theory change: Revision upon warrant. In Ph. Besnard, S. Doutre, and A. Hunter, editors, *Proceedings of the 2nd International Conference on Computational Models of Argument (COMMA 2008)*, pages 336–347, Toulouse, France, 2008. IOS Press.

C. Sakama, M. Caminada, and A. Herzig. A logical account of lying. In M. Fisher, W. van der Hoek, B. Konev, and A. Lisitsa, editors, *Logics in Artifical Intelligence*, pages 286–299. Springer, 2010.

E. Sklar, S. Parsons, and M. Davies. When is it ok to lie? a simple model of contradiction in agent-based dialogues. In I. Rahwan, P. Moraitis, and C. Reed, editors, *Proceedings of the First International Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2004)*, Lecture Notes in Computer Science, pages 251–250, New York, NY, USA, 2005. Springer.

M. Snaith and C. Reed. TOAST: online ASPIC+ implementation. In B. Verheij, S. Szeider, and S. Woltran, editors, *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*, pages 509–510, Vienna, Austria, 2012. IOS Press.

M. Snaith, J. Lawrence, and C. Reed. Mixed initiative argument in public deliberation. In F. De Cindo, A. Macintosh, and C.Peraboni, editors, *From e-Participation to Online Deliberation: Proceedings of the fourth international conference on Online Deliberation (OD2010)*, Leeds, UK, 2010.

M. South, G. Vreeswijk, and J. Fox. Dungine: a Java Dung reasoner. In Ph. Besnard, S. Doutre, and A. Hunter, editors, *Proceedings of the 2nd International Conference on Computational Models of Argument (COMMA 2008)*, pages 360–368, Toulouse, France, 2008. IOS Press.

T. van Gelder. The rationale for rationale. *Law, Probability and Risk*, 6(1–4):23–42, 2007.

B. Verheij, S. Szeider, and S. Woltran, editors. *Computational Models of Argument: Proceedings of COMMA 2012*, Vienna, Austria, 2012. IOS Press.

G. A. Vreeswijk. Abstract argumentation systems. *Artificial Intelligence*, 90:225–279, 1997.

G. A. W. Vreeswijk. An algorithm to compute minimally grounded and admissible defence sets in argument systems. In P.E. Dunne and T.J.M. Bench-Capon, editors, *Proceedings of the 1st International Conference on Computational Models of Argument (COMMA 2006)*, pages 109–120, Liverpool, UK, 2006. IOS Press.

D. Walton. *The New Dialectic*. University of Toronto Press, 1998.

D.N. Walton and E.C.W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New York Press, New York, 1995.

T.L. van der Weide and F. Dignum. Reasoning about and discussing preferences between arguments. In P. McBurney, S. Parsons, and I. Rahwan, editors, *Proceedings of the Eighth International workshop on Argumentation in Multi-Agent Systems (ArgMAS 2011)*, pages 117–135, Taipei, Taiwan, 2011. Springer.

M. Wooldridge. *An Introduction to MultiAgent Systems*. Wiley, 2001.

D. Zhang, N. Foo, T. Meyer, and R. Kwok. Negotiation as mutual belief revision. In D. L. McGuinness and G. Ferguson, editors, *Proceedings of the Ninteenth National Conference on Artificial Intelligence (AAAI-04)*, pages 317–322, San Jose, California, 2004. AAAI Press/The MIT Press.